

DOCUMENT RESUME

ED 414 908

IR 056 766

AUTHOR Gilchrist, Alan, Ed.
TITLE From Classification to "Knowledge Organization": Dorking Revisited or "Past is Prelude." FID Occasional Paper No. 14.
INSTITUTION International Federation for Information and Documentation, The Hague (Netherlands).
REPORT NO FID-714
ISBN ISBN-92-66-00-714-5
PUB DATE 1997-00-00
NOTE 191p.; A collection of reprints to commemorate the forty year span between the Dorking Conference (First International Study Conference on Classification Research 1957) and the Sixth International Study Conference on Classification Research (London, UK) 1997.
PUB TYPE Collected Works - General (020) -- Reports - Evaluative (142)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Abstracting; *Access to Information; Cataloging; *Classification; Computers; Documentation; Foreign Countries; Identification; Indexing; Information Processing; *Information Retrieval; Information Seeking; *Information Systems; *Knowledge Representation; Library Technical Processes; Natural Language Processing; Programming Languages; Relevance (Information Retrieval); Search Strategies; Subject Index Terms; User Needs (Information)
IDENTIFIERS Netherlands

ABSTRACT

This set of papers offers insights into some of the major developments in the field of classification and knowledge organization, and highlights many of the fundamental changes in views and theories which have taken place during the last 40 years. This document begins with a series of reminiscences from former delegates of the first International Study Conference on Classification Research which took place in Dorking, United Kingdom in 1957, and continues with a collection of 15 papers by classification specialists: "The Need for a Faceted Classification as the Basis of all Methods of Information Retrieval"; "Classification in Information Retrieval: The Twenty Years Following Dorking" (E. J. Coates); "Structure and Function in Retrieval Languages" (B. C. Vickery); "Knowledge Representation: A Brief Review" (B. C. Vickery); "Natural Language Processing for Information Retrieval" (David D. Lewis and Karen Jones Sparck); "The Testing of Index Language Devices" (Cyril W. Cleverdon and J. Mills); "Indexing and Retrieval Performance: The Logical Evidence" (Dagobert Soergel); "Reflections on TREC" (Karen Sparck Jones); "On Information Science" (Carl Keren); "Brief Communication: A Note About Information Science Research" (Gerard Salton); "Unanswered Questions in the Design of Controlled Vocabularies" (Elaine Svenonius); "Needs for Research in Indexing" (Jessica L. Milstead); "Intelligent Text Processing, and Intelligence Tradecraft" (Michael L. Weiner and Elizabeth D. Liddy); "Advanced Searching: Tricks of the Trade" (Peggy Zorn, Mary Emanoil, Lucy Marshall, and Mary Panek); and "What do People Want from Information Retrieval?" (W. Bruce Croft). (SWC)

FROM CLASSIFICATION TO "KNOWLEDGE ORGANIZATION" DORKING REVISITED OR "PAST IS PRELUDE"

*A collection of reprints to commemorate the forty year span between the Dorking Conference
(First International Study Conference on Classification Research 1957) and the Sixth
International Study Conference on Classification Research (London, UK) 1997*

Edited by

Alan Gilchrist

FID 714



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☐ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to
improve reproduction quality.

- Points of view or opinions stated in this
document do not necessarily represent
official DERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
IN OTHER THAN PAPER COPY
HAS BEEN GRANTED BY

Myra Verkuyl

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1997

International Federation for Information and Documentation (FID)
The Hague, Netherlands

BEST COPY AVAILABLE

FID is the leading international professional organization of institutions and individuals that promotes the idea that information is a critical resource that is needed in all levels of society (international, regional, national and sub-national) to:

1. improve competitiveness of business, industry and national economies,
2. advance the frontiers of science and technology,
3. strengthen possibilities for development and enhance the quality of life wherever possible,
4. improve the ability of decision-makers to make appropriate decisions,
5. stimulate educational strategies and life-long learning,
6. make expression possible in the arts and humanities.

FID Publication Number 714

FID Occasional Paper No. 14

- No. 1 From Informal Gatekeeper to Information Counselor: Emergence of a New Professional Role
- No. 2 National Information Policies
- No. 3 Management Ratios of Industries: Information Services
- No. 4 The State of the Modern Information Professional 1992-1993
- No. 5 Guide to the Use of UDC (Revised Edition: 1995)
- No. 6 Information Management for Business (Reprinted 1995)
- No. 7 Training for Information Resources Management: Section A (Tutorial on Information Resources Management)
- No. 8 Training for Information Resources Management: Section B (Case Studies)
- No. 9 Marketing Library Services: A Nuts-and-Bolts Approach (Fourth Edition, 1995)
- No. 10 Quality Management issues: A select bibliography for Library and Information Services Managers
- No. 11 Towards the Global Information Superhighway: A Non-Technical Primer for Policy Makers
- No. 12 New Roles and Challenges for Information Professionals in the Business Environment
- No. 13 ISO 9000 for Libraries and Information Centres: A Guide
- No. 14 From Classification to "Knowledge Organization": Dorking Revisited or "Past is Prelude"

Published and distributed by:

FID
P.O. Box 90402
2509 LK The Hague
Netherlands

Copyright © 1997 by the Fédération Internationale d'Information et de Documentation (FID)

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher.

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. The views expressed in this publication are those of the authors and do not necessarily reflect the opinion of FID.

ISBN 92 66 00 714 5

PRINTED IN THE NETHERLANDS

FROM CLASSIFICATION TO "KNOWLEDGE ORGANIZATION" DORKING REVISITED OR "PAST IS PRELUDE"

*A collection of reprints to commemorate the forty year span between the Dorking Conference
(First International Study Conference on Classification Research 1957) and the Sixth
International Study Conference on Classification Research (London, UK) 1997*

Edited by

Alan Gilchrist

FID 714



1997

International Federation for Information and Documentation (FID)
The Hague, Netherlands

Copyright © 1997 by the Fédération Internationale d'Information et de Documentation (FID)

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher.

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. The views expressed in this publication are those of the authors and do not necessarily reflect the opinion of FID.

ISBN 92 66 00 714 5

CONTENTS

INTRODUCTION

By Alan Gilchrist	iv
-------------------------	----

1. THE DORKING LEGACY

1.1 Contributions from delegates	
Robert Fairthorne	v
Jean Aitchison	vi
Cyril Cleverdon	vii
Eric Coates	viii
Douglas Foskett	ix
Derek Langridge	x
Jack Mills	xi
Brian Vickery	xii
Magda Whitrow	xiii
1.2 The need for a faceted classification as the basis of all methods of information retrieval	1
1.3 Coates, E.J. Classification in information retrieval: the twenty years following Dorking	11

2. NEW TOOLS

2.1 Vickery, B.C. Structure and function in retrieval languages	23
2.2 Vickery, B.C. Knowledge representation: a brief review	35
2.3 Lewis, David D. and Sparck Jones, Karen. Natural language processing for information retrieval	49

3. TESTING...TESTING...

3.1 Cleverdon, Cyril W. and Mills J. The testing of index language devices	65
3.2 Soergel, Dagobert. Indexing and retrieval performance: the logical evidence	85
3.3 Sparck Jones, Karen. Reflections on TREC	101

4. QUESTIONS AND ANSWERS

4.1 Keren, Carl. On information science	129
4.2 Salton, Gerard. Brief Communication. A note about information science research	131
4.3 Svenonius, Elaine. Unanswered questions in the design of controlled vocabularies	137
4.4 Milstead, Jessica L. Needs for research in indexing	151
4.5 Weiner, Michael L. and Liddy, Elizabeth D. Intelligent text processing, and intelligence tradecraft	161
4.6 Zorn, Peggy <i>et al</i> : Advanced searching: tricks of the trade	169
4.7 Croft, W. Bruce. What do people want from information retrieval?	181

INTRODUCTION

The 6th International Study Conference on Classification Research to be held in London, UK in 1997 brings this event back to the UK after meetings in Elsinore, Denmark (1964); Bombay, India (1975); Augsburg, Germany (1982) and Toronto, Canada (1991). Consequently, the organisers of the 1997 Conference thought it appropriate that the London meeting should both remember the seminal Dorking Conference and look forward; at a time when, to use the term coined by Ravetz, we appear to be experiencing a "paradigm shift" in the whole area of information handling.

In one respect, Dorking appears to belong to another more measured and scholarly age. One of the delegates was Dr. S.R. Ranganathan the eminent mathematician and librarian, who had first propounded his theories of facet analysis in the 1930s (part inspired, it is said, by seeing a Meccano set at Hamley's toy shop in Regent Street).

The Conference dinner was addressed by W. Berwick Sayers, the classification theorist (then in his eighties) whose classical work on the subject has recently been reissued for the fifth time in a revised edition.

On the other hand, the Dorking delegates were deeply conscious of the impending revolution in information technology; and it is interesting to note, for example, that one of the delegates was Eugene Garfield who later created the Science Citation Index, the foundation stone of the commercially successful Institute for Scientific Information. But Gene was not the only delegate to go on to produce significant contributions to library and information science, and it is a great pleasure to be able to catch some of the flavour of Dorking and subsequent achievements (largely through the focus of the Classification Research Group) in the contributions from some of "those who were there" that follow this Introduction.

What follows is not a comprehensive survey of the intervening years between the Dorking 'Classification for Information Retrieval' and the London 'Knowledge Organization for Information Retrieval'. Rather, it is a snapshot tribute to the Dorking pioneers, which recognizes the links between Dorking and London and that information retrieval is still a topic worthy of scholarly attention.

Alan Gilchrist

LETTER FROM R.A. FAIRTHORNE

Dear Alan,

Thanks for your letter of 3rd June and rather belatedly for your Xmas card and chronological analysis. In poor return I enclose a copy of my contribution to the USA Bicentennial (1976), which may amuse you. Also I thank you for enquiry about my health. I have had no more falls but my legs are increasingly unreliable and I begin to doubt whether the human anatomy is designed to last for ever. But I look after myself and expect to go on that way beyond my 92nd birthday in a week.

You suggest that a frog or frogs might contribute. Unfortunately, the frogs did not evolve till many years after Dorking. They were generated from the confluence of lengthy committee meetings, blank spaces on agenda papers, and the invulnerability of the batrachian outline to poor draftsmanship. However, should you welcome an invasion by anachronistic frogs, I enclose two copies of a possible contribution. For their comment the frogs have borrowed from a Federal building in Washington D.C.

After forty years my recollections of Dorking may not be worth much, but one aspect remains intact; to wit, how pleasant it was. I recall, in an after dinner get together, Jesse Shera's praise of the BBC's programme about "The Spaghetti Harvest in Italy"¹ leading to an informed discussion on the Goon Shows. Also a presentation to myself, for no discernible reason, of a home made Flexohedron. In particular, Gene Garfield's surprise at British members disagreeing without being disagreeable. This verdict was based on far too small a sample but, in spite of statistical criteria, was valid for all nationalities at this and subsequent conferences.

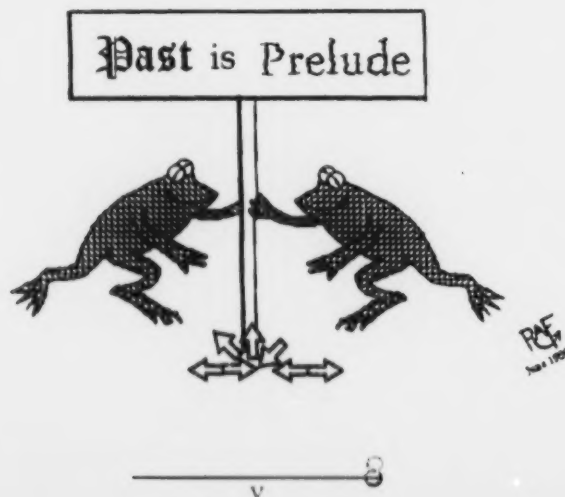
Dorking was necessary, if only to show that devices by themselves were not enough to achieve inverse communication. It was successful because it set in motion the understanding that classification is an essential ingredient of informational activity.

Sincerely

R.A. Fairthorne

NOTE

1. On April Fool's Day the BBC broadcast a short film showing farmers cutting down the long strands of spaghetti, accompanied by a poker-faced commentary by (if I remember correctly) Richard Dimbelby (ed.)



THE TIME OF DORKING

BY CYRIL CLEVERDON

The early 1950s were a time of general ferment in the area of what is now generally known as information retrieval. This had been stimulated by the great increase in scientific and technical information resulting from the lifting of security restrictions on the tens of thousands of research reports published during the war, and the consequent feeling of many scientists that the conventional techniques of indexing could not cope.

In Europe generally, and in Britain in particular, the most common tool was the Universal Decimal Classification (UDC), but it was sadly out-of-date. Under the umbrella of the British Standards Institution, several subject groups were set up to amend the schedules. However, the most exciting work was coming from the Classification Research Group (CRG), a small group of a dozen or so individuals who met monthly to develop the theory of facet classification, as originally propounded by Professor Ranganathan. Continuing the theme of classification was the work undertaken by Jack Mills to revise the Bliss scheme.

In spite of it being the home of Dewey, whose classification was the basis of the UDC, in the United States the indexing of S&T literature had been by alphabetical subject catalogues, and the first person seriously to challenge this method was Mortimer Taube with his Unitary System of Coordinate Indexing (or, as we would now say, 'post-coordinate indexing') as opposed to the pre-coordinate indexing of an alphabetical subject catalogue. Others proposing new techniques were Calvin Mooets with Zator-coding, John Berry with semantic factoring, and, perhaps most revolutionary of all, citation indexing as proposed by Gene Garfield. (On a personal note, my most vivid memory of Dorking is an evening in the bar when Gene defended his proposals against a group of very sceptical and outspoken critics)

The Aslib Aeronautical Group (A.A.G) had, from its inception in 1950, been keenly interested in indexing techniques and its 1955 Cranfield Conference was a joint meeting with the C.R.G. From that conference came the idea of holding a conference dedicated to indexing systems, and in particular to classification. Also was born the idea of the A.A.G developing a comparative test of the efficiency of conventional systems (alphabetical and UDC) and the new systems (Facet and Uniterm), for which the CRG undertook to develop a special Facet classification. A proposal was developed and by the time of the Dorking Conference, two members of the CRG, Brian Vickery and Jason Farradane were developing a Facet Classification of aeronautical literature. As it turned out there was little significant difference between the efficiency of the old and new systems of indexing.

At the time of Dorking, the use of computers in information retrieval was only an idea, and it was impossible to foresee the increase in computer power which would permit the analysis of full text using complex statistical techniques for matching queries and documents. However, in spite of thirty years development along these lines, the gain in performance, as measured by recall and precision, appears to be slight. Perhaps a bringing together of the logic of classification and the power of modern computers could be a way forward. Maybe the 1997 Conference will give a lead.

However, in doing this one hopes that the new generation of researchers will take note of the lessons of the past, for, to quote Dr. Michael Lesk, 'Much of what was done in the 1960s was done so well that it does not need to be done again - even though it often is'.

RECOLLECTIONS OF THE DORKING CONFERENCE - MAY 1957

BY JEAN AITCHISON (Née Binns)

Back in the May 1957, before thesauri, before any significant application of computers in information systems, when 'knowledge organization' was known as plain 'classification' and full-text searching was hardly imaginable, I attended the Dorking Conference. I was there because I was already a member of the Classification Research Group, bitten by the faceted classification 'bug', and about to start work on the English Electric Faceted Classification for Engineering, which eventually led to Thesaurofacet.

My first memory is of conference members assembled at Dorking railway station, waiting for a bus to take them through the beautiful Surrey countryside to the conference centre at Beatrice Webb House, set among trees on a hillside. I mention in my notes that 'the catering at the first meal was institutional, but the conversation was bright', not surprising, considering the concentration in the dining-room of legendary 'greats' from seven countries, such as S.R. Ranganathan, M.G. Cordonnier, Eric de Grolier, Cyril Cleverdon, Brian Vickery, Douglas Foskett, Jack Mills and many others.

From memory and notes, I can recall some of the papers, in particular that of S.R. Ranganathan, demonstrating the Colon system; Jack Mills lucidly explaining subject-field analysis, Eric Coates advocating non-structural notation, and the brilliant concluding survey by de Grolier. A vivid memory is of Eugene Garfield, as a young man, vigorously marketing his idea of journal contents lists, at an extra evening session. I also recollect W.A. Wildhack chairing, with flair, the last evening's session, when Robert Fairthorne entertained us with his famous 'paper tearing' act, making bow ties out of dollar bills and ten-shilling notes.

During the drafting of the conference recommendations, with Douglas Foskett in the chair, I was surprised at the degree of agreement, and especially at the ease with which the recommendation emphasising the importance of categorised (or faceted) classification was adopted.

The Cranfield studies, which followed soon after Dorking, and subsequent research, have shown that classification is only one among many factors affecting retrieval performance. However, facet analysis, if not the panacea the Dorking recommendations considered it to be, has still a significant role to play in information retrieval. This may be seen in the number of papers on facet techniques still appearing in the journal, *Knowledge Organization*, and at ISKO and other conferences, and in its continued influence on thesaurus construction and on the development of international classification systems such as the Bliss Bibliographic Classification (BC2) and the UDC.

1957-1997 *PLUS ÇA CHANGE...* WHAT HAS HAPPENED TO CLASSIFICATION SINCE THE DORKING CONFERENCE?

BY E.J. COATES

A minor but perhaps not insignificant revelation which surfaced during the 1st International Study Conference on Classification for Information Retrieval at Dorking in 1957 was that taxonomists (and by implication other kinds of compilers of classification systems) were engaged both in the "splitting" and "lumping" of concepts. Young splitters, we learned, tended to become lumpers with advancing age. For an elderly lumper simultaneously looking back at the Dorking meeting and forward to the proposed 6th International Study Conference being planned for 1997 there is a mixed sense both of *déjà vu* and of immense change. Forty years on finds a world dominated by costs against benefits which often cannot convincingly be measured. At the 1957 meeting costs were explicitly mentioned once only, by Ranganathan, who ventured to declare that "Machinery can do the work at great speed; but it is more costly, and it requires a vast quantity of minimum turn-out to become economical".

In the intervening years the cost aspects as between human and machine handling of information work have, of course, been dramatically reversed, concomitantly with a vast increase in potentially accessible information resources. Associated with this cost reversal there has been a decline, in both quantitative and qualitative senses, in the application of classification to libraries and information retrieval generally. Not that the Dorking conference turned out to be a sterile talking shop. It was followed by a plethora of practical systems and projects which reflected its influence. One had to think only of such examples as the special subject classifications created by members of the Classification Research Group, the wider angled faceted schemes produced by Jean Aitchison, the BTI indexing system, PRECIS, the new Bliss Classification (BC2) and the Broad System of Ordering. Despite their being successful demonstrations of faceted classification principles in action, interest in them faded in a climate dominated by mechanisation and short-term cost savings on human intellectual work. In one respect mechanisation was beneficial in clearing the air, in exposing the erroneous claim apparently endorsed by the Dorking conference that faceted classification was a necessity for *all* methods of information retrieval. Some, but not all, subject retrieval requirements are in fact adequately satisfied by computer scanned word matching. What computers cannot reliably do without having recourse to the most fully-structured available classification systems, is to generate appropriate concept tags summarising a given unit of natural language text.

It is for this reason, among others, that despite the undoubted achievements of computerised retrieval, human intermediaries interacting between searchers and systems are often needed, and that there is a certain undercurrent of dissatisfaction with what is at present on offer. Nowhere is this more apparent than on the Internet, where some users are demanding that such a vast array of knowledge should be better organised if it is to become easily accessible. That this instinct is soundly based finds ample confirmation in the papers of Shera and Ball presented at Dorking. There are signs that Internet culture is groping, albeit ponderously, in the direction to which 20th century classification thinking has been pointing. Some search engine suppliers are beginning to talk approvingly of concept indexing, though their practical steps to this end (in so far as commercial secrecy permits them to be visible at all) seem to be highly dependent upon subjective decisions taken against inadequately structured background understanding of classification. The Dorking meeting was strong on facets and syntactic relations, but had almost nothing to say about the semantic axis of classification. It is perhaps the time to redress this lack of balance at the 1997 conference. Pursuit of this relatively neglected line of development would inevitably lead into the realm of user behaviour in subject searching, and beyond that to the social and cultural environments affecting classification - those same environment factors which have contributed to the temporary eclipse of classification for retrieval in the last two decades.

"DORKING REVISITED"

BY D.J. FOSKETT

Bliss was it in that dawn to be alive!" We might well have paraphrased Wordsworth's cry, "Bliss in that dawn became alive!" because the renaissance of H.E. Bliss's great Bibliographic Classification could be said to have begun at "Dorking '57". The main impetus came from the Classification Research Group, which recently held its 304th meeting, at University College London, and can point to many volumes of new editions of that scheme.

"Dorking '57" was unusual, possibly unique at the time, in assembling people from various disciplines and countries to discuss classification and its role in information storage and retrieval. A magisterial welcome was given by W.C. Berwick Sayers, then nearly 80, to this gathering, he said, of "retrieval masters".

Some important names were missing, most notably, I think, Hans Peter Luhn, whose presence might have secured the interest of IBM. Some present have taken little part in subsequent work, but there was a fruitful and constructive exchange of views from many fields.

Good groundwork was done by the CRG, especially by Brian Vickery, and the course was firmly set in the direction of facet analysis for all forms of retrieval systems, now embodied in the Bliss BC revisions. The late Bernard Palmer, who did as much as anyone to introduce S.R. Ranganathan to the West, particularly our colleagues from the USA, was an urbane and capable Chairman. Ranganathan's presence and lecture began a memorable series of arguments with Eric de Grolier, whose concluding summary was a foretaste of many more brilliant analyses of conference proceedings.

"Dorking '57" was an outstanding event in LIS history and in the lives of those attending, who remember it with advantage and affection. May "Dorking revisited" be as important and as fortunate.

A NEW DAWN? BY D.W. LANGRIDGE

Bliss was it in that dawn to be alive". Thus an enthusiast for the French Revolution, who might well have been echoed by an enthusiast for classification at the time of Dorking. We were confident then about the importance of classification and about its future prospects. Disillusionment of enthusiasts for the French Revolution was swift: disillusionment of enthusiasts for classification was slower in coming but no less sure.

The Dorking proceedings opened with Ranganathan on "Classification as a discipline" and closed with the C.R.G.'s credo "The need for a faceted classification as the basis for all methods of information retrieval." By the end of the sixties this latter idea had been swept away by the misguided pre-occupation with computers. The steady flow of books on classification theory and practice dried up. Organisations that badly needed more effective subject retrieval preferred spending money on computerising their catalogues to improving their classification. The British Library, who had the opportunity to take the lead in sponsoring a modern general classification when its new building was first planned, chose instead to remain a hundred years behind the times. C.I.P. was a good idea, badly executed. Only a few people excel at classification and we needed above all a central body providing *authoritative* subject analysis. This we did not get from C.I.P.: another fine opportunity wasted. Provision of a modern general classification has been left to the Bliss Society with minimal financial support, and apathy from the profession in general.

Since bibliographic classification is essentially a practical discipline these developments are extremely discouraging to anyone devoted to classification research. I think we can only have faith that the last 25 years will soon be recognised as an aberration, and continue to make whatever contribution we can - even if it doesn't bear fruit in our own lifetime.

Classification as a discipline has also suffered. Its central importance to librarianship has been forgotten and its teaching in library schools drastically reduced. Among those concerned with the organisation of information there has been a proliferation of methods and terminology. Dorking proceedings were reasonably unified: we all understood each other. The same can not be said for the proceedings of subsequent conferences. There are, however, some optimistic signs and perhaps the forthcoming conference will herald a new dawn.

COMMENTS ON DORKING AND AFTER

BY JACK MILLS

I think of Dorking as a highlight in the recognition by librarians of the importance of classification in their everyday work - but one which proved to be something of a false dawn. Libraries were still largely using antiquated 19th century type classifications as their central management tool for organizing the information they stored. The Classification Research Group (CRG) seemed to me to represent a movement which might provide a lever whereby the profession could be moved out of the torpor of unthinking pragmatism which this addiction reflected.

Jesse Shera (a Dorking participant) had already made his masterly summary of what running a library was about. Only two things, really: bibliography (his shorthand for knowing about, selecting and managing the materials constituting the store) and information retrieval (his shorthand for helping the users find what they need, faced with the millions of potential places there are to look in). It could be argued that librarians did appreciate that classification was the basis of retrieval, which always boils down to the separating out of a tiny subset or class (however defined) of items from the store. Evidence of this was the willingness of libraries to spend millions of dollars (it was usually dollars) in reclassification. But this was mainly to change from DC to LC. Imagine - from 1876 to 1897!

But here were some librarians prepared to develop better management tools for the organization of their store and of the ever-increasing flood of information. Surely the profession was beginning to move at last? No hope: what was moving was the burgeoning flood of research into an IR which was equated with one particular mode of retrieval - that of specific bits of information, from largely scientific and technical information stores. Simple locating was overwhelmingly predominant as compared with relating. But the two operations - locating and relating - have long been recognized by librarians as the essence of all indexing (using this term in its broadest sense of indicating where to begin looking and where else to look, assuming that the first simple locating is insufficient, as in practice it usually is).

To compound this side-tracking of the new dawn for information-classification, the computer was advancing rapidly and was soon to be seen in the popular mind as the IR machine *par excellence*. Its awesome capabilities, combined with its steadily diminishing cost, reinforced the myth that the cost of IR itself could be drastically reduced by throwing overboard the expensive, labour-intensive study of knowledge structure and relying on simplistic procedures which a machine can handle. These procedures centre on the simple coordination of keywords with minimal control of relationships (which is 'expensive').

To my mind, this view of the economy of information storage and retrieval (*aka* library management) has unhappy echoes of the economic fashions of the last two decades. These have favoured measures of performance and 'efficiency' which are made within the framework or accounting systems whose boundaries are drawn unfairly, with the omission of many important factors and a premium put on short-termism. But, as in the case of these fashions in economics, a more sensible policy, based on a wider perspective, will, it is hoped, emerge in IR. That the bedrock of any information system is the structure of knowledge seems unarguable; and classification is the key to its comprehension and its practical management. The spirit of Dorking lives on!

LOOKING BACK...

BY BRIAN VICKERY

Looking back at the proceedings of the Dorking conference, I find that I was reported as saying that "...The theory of information is today in a state of rapid development. Many new schools of thought are forming - often in polemic opposition to each other. We have much to learn from other disciplines - from mathematics, from communication theory, from data processing, and from linguistics. We have a long way to go before the contributions from all these fields are properly integrated into a general theory of information."

The same statement could be made at any subsequent period since Dorking. The emergence of cognitive science has brought about a certain amount of unification, but we still have 'a long way to go'. In the 1950s classification was regarded as a key problem in information retrieval, but was rarely considered important in other disciplines. Its status in retrieval has waned and waxed over the years, but increasingly it has begun to play a part in wider areas of study. In 1962 a paper on 'classification for documentation'¹ was prepared as a contribution to a meeting that looked at classification in several fields, in particular numerical taxonomy. By the 1970s, the use of hierarchical semantic nets and of 'roles' equivalent to information's facets, became common in studies of human information processing and in artificial intelligence.² In the 1980s, categories and classes were being used in intelligent online search interfaces.³ Now in the 1990s the new techniques of knowledge discovery in databases find that automatic classification and clustering algorithms are of central importance.⁴

In his concluding address at Dorking, Eric de Grolier closed the conference by asserting: 'If I had a glass of champagne, I would lift my cup to the future classificationist who will be master of the world!' Well, we all have occasional illusions of grandeur, but it seems certain that classification will continue to play a part in many aspects of cognitive science and its applications, and in itself will be one of the unifying factors.

NOTES

1. Vickery, B.C. Classification for documentation. *Aslib Proceedings*, 14, 1962, 243-7
2. Vickery, B.C. Knowledge representation: a brief review. *Journal of Documentation*, 42, 1986, 145-59
3. Vickery, B.C. Categories and classes in search interfaces. In: *Libraries and Information services. Education Libraries Journal Supplement* number 25, 1993
4. Frawley, W.J. et al. Knowledge discovery in databases: an overview. *AI Magazine*, Fall 1992, 57-70

DORKING REVISITED

BY MAGDA WHITROW

In a concluding survey of the problems of classification for information retrieval at the conclusion of the International Conference on Classification, held at the Beatrice Webb House, near Dorking from 13th to 17th May 1957, Eric de Grolier began by saying that "We are approximately at the beginning of a truly scientific study of classification", but he warned that to prepare a general encyclopaedic scheme of classification we must wait one or two years before being able to determine exactly on what basis such a scheme must be constructed.

That was forty years ago. Looking through the Proceedings of the Conference it seems that in fact that basis was already laid down at that meeting and the schemes produced by those members of the Classification Research Group who were present then and are still active members of the Group have made significant contributions to the theory and practice of classification. I am referring to Jean Aitchison, Eric Coates, Douglas Foskett, Derek Langridge and last but not least to Jack Mills whose monumental work on the revision of the Bliss Classification is outstanding.

I joined the Classification Research Group some time after the Conference which I remember with some nostalgia. It was a very happy meeting - there were only forty members and the surroundings were very beautiful. I would like to take this opportunity of paying tribute to some of those present then who made important contributions to classificatory theory and practice, but are no longer with us: D.J. Campbell, Jason Farradane, Jack Wells and in particular Barbara Kyle who was the life and soul of the Conference and who died tragically early. I owe her a great debt of gratitude. It was she who encouraged me in my attempts to develop a scheme of classification of a special subject and prepare it for publication.

THE NEED FOR A FACETED CLASSIFICATION AS THE BASIS OF ALL METHODS OF INFORMATION RETRIEVAL

Memorandum of the Classification Research Group

Received, for information, by the L.A. Library Research Committee,
May, 1955

INTRODUCTION

During the past three years, a Classification Research Group has been meeting in London. It was formed as a result of a Scientific Information Conference held in 1948. The Royal Society of London at that time set up a standing committee on subject classification in science. At the end of 1950 this committee reported that its discussions "have served to show that the problem of classification is far more complex than was imagined at the outset". Its secretary, Professor J. D. Bernal, suggested a study of the problem by interested librarians. After many months of informal discussion, the latter agreed, in February, 1952, to form a Classification Research Group, whose present members are listed in an appendix.

The aim of the Group has been to review the basic principles of bibliographical classification, unhampered by allegiance to any particular published scheme. It has advanced slowly, carefully recording the points upon which agreement has been reached, and has so far published one brief, corporate statement of its views (RECORD, June, 1953, pp. 187-188).

The first meeting of the Unesco Advisory Committee for Documentation in the Natural Sciences was held in Paris, 15th-17th February, 1954, and considered Recommendation No. 19 of the International Conference on Science Abstracting, 1949. This was to the effect "that ... Unesco support the development of a standard classification", and that Unesco should facilitate the production of a standardized international code for automatic selection. The Advisory Committee "deprecates assuming *a priori* that the U.D.C. or any other established classification is the best method of coding" (Unesco document NS/115/4405, dated 9th April, 1954). The Unesco *Monthly circular on documentation and terminology in the natural sciences* (1954-04) announced that a series of reports and memoranda would be issued on "methods of literature referencing and of retrospective searching". The present paper is written as a contribution to that series.

SUMMARY

It will be as well to set out at once, in summary fashion, the point of view here presented. An index, a classification, an automatic selector, or any other system of "information retrieval", is a working tool designed to help the user to find his way about the mass of published information relating to a certain field of knowledge. The user may have a detailed understanding of the pattern of knowledge in the subject he explores, or he may have only an uncertain and confused understanding of it. An information retrieval system should be designed (1) to help even the ignorant user to pass from the vague formulation of a subject in his mind to its precise formulation in the system, and (2) having reached this precise formulation, to direct the searcher forward to literature references relating to it.

The paper which follows examines four types of retrieval system: the alphabetical subject index, the co-ordinate indexing system, the classified index, and automatic selection.

It is clear that all four are equally able to serve the second purpose stated in the last paragraph though they may differ in ease and speed of use, and in ease and speed of construction. The paper then goes on to consider how these

four systems may be designed so as to provide help to the less knowledgeable user. Put briefly, the conclusion to which we are drawn is that this can be done only if each system is based on a classified schedule. In the alphabetical and co-ordinate indexes, this schedule must serve as the source of unit terms and the basis whereby appropriate links between related terms are forged in the index. In the coding for automatic selection, and much more fully in the notation of a classified index, the schedule is the framework which directly displays the linkage of related terms.

The most important relation which the schedule displays is the class relation, between a generic term and its collateral species. This paper examines whether the generic relations of the field of knowledge can be adequately displayed by a simple enumerative tree-of-knowledge classification. It concludes that they cannot. The terms subsumed under a given genus - e.g., a main class - are not all derived from that genus by differentiation using a single characteristic or division. They can be sorted into groups or facets, each of which is derived from the genus by a different characteristic. In order adequately to display the linkage of related terms, it is necessary to have faceted classification schedules.

The main interest of the Classification Research Group is to study in more detail the ways by which such schedules may be constructed. The study is a continuing process which is far from finality. The later parts of the paper discuss briefly two problems to which attention has been directed: techniques for facet-analysis, the sorting of terms into groups or categories, and notational devices which may be used in a classified index.

The views outlined above will now be set out in more detail.

INFORMATION RETRIEVAL

The process of acquiring documentary information is not a simple one, and can be analysed into several stages:

- (1) Identifying the exact subject of search.
- (2) Locating this subject in a guide which refers the searcher to one or more documents.
- (3) Locating the documents.
- (4) Locating the required information in the documents.

The first stage is the process whereby the original formulation of a subject in the searcher's mind is transformed into the particular symbolism used in the guide of stage 2. The original formulation may exactly represent the subject sought, but be in a form unsuited to the guide (e.g., in words instead of class or code numbers); or it may not correctly represent the subject - the inability even of specialist enquirers to state their requirements exactly is well known to librarians.

The second stage may be multiple. For example, consider a search for information on "diquinolalophenazine", using as a guide the index to *Chemical Abstracts* for 1926. The index directs the searcher onwards, by means of a page number, to a second guide, the abstracts themselves. This guide must then be used to locate a reference, *J. Chem. Soc.*, 1926, 1171-84. The third step, locating this document, involves the use of further guides (union catalogues, library shelf-lists), and the fourth step is to study the document. The fourteen pages of the article must at least be scanned, and the compound sought be recognized under the name of "phenazineazineazine".

Three mechanisms of "information retrieval" are generally recognized - indexing, classification and automatic selection - and Taube has recently added a fourth co-ordinate indexing. Some writers consider these techniques to be alternative ways of carrying out the same process, differing only in scope and effectiveness. Others consider them to be mutually exclusive, each suitable only for one aspect of information retrieval. It is necessary to examine how each fits into the pattern of search set out above.

The four mechanisms are not concerned with the last two stages of the search. Stage 3 is concerned with locating a document by means of author or title lists, not by subject, and stage 4 is the detailed study of a document. It is only the first two stages which concern us. Let us consider them more closely.

Stage 2 is relatively simple. Having discovered, in the first step, the appropriate symbolism by which the exact subject sought is represented in the guide, the latter is interrogated to ascertain the literature references it contains which are relevant to the subject. The appropriate symbolism may be a subject heading of one or more words, a class number (simple or compound), or one or more code numbers. With these, the "guide" - alphabetical index, classified index, co-ordinate index or automatic selector - is interrogated in whatever way is appropriate to its structure. At this stage, therefore, the four mechanisms are completely equivalent, assuming that they have all been constructed to contain the same volume and detail of information. They may differ in ease and speed of construction, or in ease and speed of use, but they perform the same function and achieve the same result. Therefore, as far as stage 2 of information retrieval is concerned, the four mechanisms can only be evaluated economically - as to the amount of work in construction and use which each requires. This evaluation is an important problem to which little time has yet been devoted.

THE STRUCTURE OF SUBJECTS

Consideration of stage 2 of information retrieval has brought out the feature which the four techniques here discussed have in common. They appear as alternative ways of carrying out the same process. A study of how they must be modified so as to cater for stage 1 will now be made.

Let us consider the nature of a specific subject, the themes on which books, parts of books, articles, or parts of articles are written. What constitutes such a subject? A study of book titles alone would suggest that literary subjects have simple names like "War", "Religion", "Boats", "Musical pitch", "Colour", "Acridines", "Wild flowers", and so on. But the study of articles at the documentation level reveals that such titles are simple in appearance only. Such a subject is in reality a complex aggregate of specific subjects - each of which is the main theme discussed from one particular aspect. To put the matter more precisely, the specific subject of an article is not a simple concept which can be neatly tucked away in a single pigeon-hole in the vast cabinet of knowledge. It is a compound, more or less complex, of simpler concepts.

This fact is recognized in all types of system designed for information retrieval. The alphabetical index uses compound subject heading such as *electromagnetic-waves-propagation-equations*. The co-ordinate indexing system of Taube lists each of these four words as a Uniterm in an alphabetical series, entering the appropriate literature references against each. In an automatic selection system, such as Mooers' Zatocoding, each of these terms or "descriptors" is separately coded and punched on a card which records a literature reference. An analytico-synthetic classification scheme, such as Ranganathan or Cordonnier have devised, transforms each constituent term into a classification symbol and represents the subject by an aggregate of such symbols. All these techniques recognize the compound nature of specific subjects.

What is needed in order to interrogate these various mechanisms, to discover references on a particular subject, such as the one cited in the previous paragraph? In other words, what must stage 1 of the retrieval process provide before stage 2 can be carried out? (a) For interrogating the co-ordinate index, the individual unit terms, which collectively describe the exact subject sought, must be known. (b) For a normal subject index, not only must the unit terms be known, but also the order in which they are cited in the compound heading. (c) For a synthetic classification, both terms and order of citation must be known, and also the symbol by which each term is represented. (d) For automatic selection, the terms must be known, also the code by which each is represented; the order of citation (in this case, the "field" of punching) may or may not be needed, according to the type of selector used.

At first glance, this analysis seems to point to co-ordinate indexing as the solution to problems of information retrieval, since its requirements for stage 2 are simpler than in the other mechanisms. But it must be recognized that this analysis is concerned only with the *amount* of information which stage 1

has to provide in order to interrogate the index in stage 2. It has not revealed how, or how readily, this information can be obtained - i.e., it has not analysed the mechanism whereby stage 1, identifying the exact subject of search, is carried out. Before we evaluate the different techniques of information retrieval, their performance at both stages of the process must be examined. The operation must be studied as a whole.

IDENTIFYING THE SUBJECT

In order to carry out stage 2 of the literature search, the literature references must be incorporated in the following mechanisms. (a) For a co-ordinate index, an alphabetical list of unit terms; (b) for a subject index, an alphabetical list of compound headings; (c) for a classified index, a serial list of class numbers; and (d) for automatic selection, a collection of coded cards (or other devices) and a sorting mechanism. What additional devices are needed in order to carry out stage 1?

(1) Even if we assume that the searcher has correctly expressed in words the subject he seeks, it is still necessary to provide an alphabetical list of unit terms to translate the words into either class or code symbols, in cases (c) and (d) above.

(2) The searcher may choose the correct *concepts* to express the subject, but may not use the standard words used in the index. He may, for example use the word *hertzian* instead of *electromagnetic*. It is thus necessary to incorporate in the alphabetical index - in all four cases considered above - cross references from synonyms. These synonyms will be chosen from dictionaries, etc., and will rarely cover all possible variants.

(3) The searcher may use the correct words to express the subject, but may not choose the correct order of citation. This problem only arises for cases (b) and (c) above. The searcher may be compared to an observer who has analysed a witnessed event into its correct elements - Tom, Bill, Alice, saw, and beat - but does not know how to combine the words so as to make clear that it was Alice who saw Tom beat Bill. The searcher is more handicapped than the observer, since the order of citation in a subject heading is dictated not only by meaning, but also by cataloguing conventions which may not be obvious to the uninitiated.

The usual device adopted in subject indexes is to use one standard order of citation (e.g., *electromagnetic-waves-propagation-equations*) and to make cross references from other possible headings. For a binary heading, there is one other possible order, and for a ternary, five. For the quaternary here cited, there are 23 theoretically possible other orders. About half of these express the meaning of the subject adequately and so should be, included as cross references. The "chain indexing" technique of Ranganathan keeps down the number of cross references to $n-1$ for a compound of n terms by listing only certain combinations.

An alternative device, which is adopted in the classified index, is to lay down rules which regulate the order of citation precisely. By constructing a class heading according to these rules, the searcher is led unequivocally to the correct order of citation. The essence of this device is the grouping of the unit terms into categories, and the establishing of rules which govern the order of citation of these categories. The result of this procedure is that, in the classified index itself, a compound subject need be entered only once - under the compound class number which uniquely represents it. In theory, therefore, this categorization would avoid the necessity of cross references. In practice, however, the construction of a class symbol is a skilled job, because the complex detail of the subject matter classified inevitably necessitates rather intricate rules of classification. The searcher cannot be expected to master these. Consequently, while only one entry is needed in the classified index, in order to help the searcher the usual cross references are needed in its alphabetical supplement. The latter must therefore include both unit terms and compound subject headings.

(4) The searcher may choose the correct terms, but may use only some of those which collectively describe the subject. He may, for example, seek the subject of section (3) under the heading *electromagnetic waves - equations*. This will not matter in the case of mechanisms (a) and (d) -

co-ordinate index and automatic selection. If the full cross references of section (3) are used, the searcher will be directed to the correct entry in a subject or classified index. The limited references produced by "chain indexing" may not lead him directly to the correct entry.

(5) Lastly, the searcher may choose concepts which do not correctly represent the subject sought. It must not be supposed that this is a rare occurrence. In a recent experimental study, Oliver Lilley analysed 368 incorrect headings which had been suggested by users to represent the subject of six books. In 79.08 per cent of these, the terminology was incorrect (i.e., a synonym was used); in 72.28 per cent, an incorrect form (i.e., order of citation) was used; but incorrect specificity (i.e., the wrong concept) was used in 93.48 per cent of the examples analysed. That is to say, of those who failed to name the subject exactly, nearly all - as well as using synonyms and incorrect order - used the wrong concepts (*Amer. Doc.*, vol. 5, no. 2, pages 41-60, April, 1954). The concepts used are not, of course, entirely unrelated to the correct concepts - they are usually broader, more general concepts which include the correct ones.

All four types of search mechanism - co-ordinate index, subject index, classified index and automatic selection - are faced with this problem. The principle which must be adopted to solve it is quite clear - a reference from each concept to those which are subordinate to it. These references can be made casually as they occur to the indexer. The only way to make them consistently is to draw the unit terms from classified schedules in which their generic relations are clearly set out. The classified index does this automatically. An alphabetical subject index with full cross-reference from each term to those which are subordinate to it can be constructed from classified schedules by means of "chain indexing". In the co-ordinate index, all the higher collectives of each term used can be drawn from a classified schedule and inserted as Unit terms into the index. An automatic selection system can indicate generic relation only to the extent that the coding can be put into the form of a hierarchical notation.

It remains to be noted, that the chosen concept may not simply be more general than the correct one - it may be related to it in a less simple way. Only one type of mechanism - the classified index - is in any position to deal with this problem. If to the serial list of numbers are attached verbal "feature" headings, expressing the structure of the classification in words, the pattern of knowledge is made visible to the searcher. He may then often be led towards the correct heading. In all the other mechanisms, the pattern of knowledge - though present in the classified schedules from which the *indexer* draws his terms - is concealed from the user.

The mechanisms needed in order to carry out the first 2 stages of information retrieval can now be summarized in the following table.

Mechanism	Coordinate	Subject	Classified	Automatic
Alphabetical list, containing				
Unit terms and synonyms	Yes	Yes	Yes	Yes
Compound headings ..	-	Yes	Yes	-
Inverted headings ..	-	Yes	Yes	-
Subordinate references ..	-	Yes	-	-
Schedule of Classes ..	Yes	Yes	Yes	Yes
Serial list of classes ..	-	-	Yes	-
Coded cards and sorter ..	-	-	-	Yes

Once again, as a result of this analysis, we find that it is the co-ordinate index which has the simplest mechanism. It needs only a list of unit terms and a classified schedule from which to draw them. The other three mechanisms need these two devices, and other devices in addition. It is clear, therefore, that the co-ordinate index is the simplest of these four mechanisms to construct. We must now consider their relative ease of use.

What are the *main* time-consuming steps in each of these four mechanisms? In the co-ordinate index and automatic selection, time is spent in physically selecting the references. In the subject index, it is in

following up cross-references from one term to the next. In the classified index, it is in "following the map" to reach the subject sought. We have no data from which to estimate the objective time spent in each case, but it seems that the time spent in working through a map of knowledge might well be "subjectively" much less than that occupied by selecting references from a co-ordinate index, waiting for an automatic selector to operate, or hopping from one item to another in an alphabetical index. Moreover, only the classified index offers the opportunity of correcting mistakes in choice of concepts other than degree of generality.

Assessing the relative value of these four mechanisms of information retrieval is therefore rather complex. In circumstances where the searcher will in most cases name his subject precisely, the simplest system may be advantageous. The less certain we can be of correct identification of the subject, the more valuable become the accessory aids for stage 1 of the retrieval process. In those cases where the mechanism must cater for all types of user and for all fields of knowledge, the classified index is likely to be the most valuable.

CLASSIFICATION SCHEDULES

Whether the classified index be used or not, the preceding argument has indicated that an essential tool in constructing *any* retrieval system is a classification of knowledge. The purpose of this classification is to provide schedules in which the most useful generic relations of each unit term are displayed.

The main interest of the Classification Research Group has been to consider how such a classification should be constructed. In a document circulated for its inaugural meeting, a number of criticisms of existing general classifications were put forward, as follows:

- (1) The schedules provide insufficient detail.
- (2) Revision and extension are too slow.
- (3) The schemes often disagree as to the correct placing of subjects.
- (4) Many classes are a heterogeneous jumble of tenuously related items.
- (5) The occurrence of subjects in more than one class is not satisfactorily dealt with.
- (6) Phenomena whose relations with existing knowledge are uncertain cannot readily be accommodated.
- (7) New classes of knowledge, which developed from new phenomena or by growth within an existing class, cannot readily be accommodated.
- (8) Relationships between subjects are distorted.
- (9) Classification schemes present a rigidly specified network or pathways leading to rigidly grouped collections of items.
- (10) Notational representation may be complex, confusing, lengthy and hard to remember.

The Group considers that, taking all such criticisms into account, there is no existing set of general schedules which is satisfactory, nor is there one which can be made satisfactory by a simple process of extension and revision. The Group has found the need for a prior discussion of principles by which classification schedules may be constructed to serve the purpose defined above.

This purpose is the display of useful generic relations. In connection with this, the criticism that relationships are not correctly displayed, but are distorted, may first be examined.

The traditional enumerative systems of classification assume that generic relations can be adequately displayed by the construction of one vast tree of knowledge, starting with "the universe", dividing into a series of main classes, and continuing such subdivision into whatever fine detail is required. It is stressed that this process of division must be logical, i.e., that each genus is differentiated into species by means of a particular characteristic of division. If this is true it follows that any one term will be divided in only one way, on the basis of only one characteristic, and all the subdivisions of any one term should be true species of the parent genus.

An examination of the schedules of the major classifications reveals that this is not true in practice. For example, subsumed under the heading "physical chemistry, chemical constitution (structure)", the Bibliographic Classification lists terms dealing with (a) the structure of atoms, (b) the structure of molecules, (c) principles of chemical reaction, (d) internal molecular changes, (e) interactions between molecules and light, etc. Similarly, in the Colon Classification, the "problem" class in chemistry includes reactions between molecules, the internal structure of molecules, molecular mixtures, procedures of investigation, and so on. The terms so grouped are not related in the way theory claims them to be - they are not collateral species differentiated from a genus by means of a single characteristic. The enumerative schedules thus fail to display correct relations between terms.

More detailed investigation shows that in some cases the various subdivisions of a particular term are all species of which the term is the genus, but they are not all *collateral* species. They can be sorted into groups, each of which has been differentiated from the genus on the basis of a *different* characteristic. For example, terms found in writings on chemistry can be sorted into terms relating to Substance, to State of Matter, to Reaction, to Operation, and to Property. These five groups of terms are all subdivisions of Chemistry, but each is derived from the genus in a different way.

Since this is true generally, it follows that a simple tree of knowledge cannot satisfactorily represent the generic relations between terms. A type of schedule is needed which allows a given genus to be subdivided in more than one way, to give several sets of subclasses, each of which is a homogeneous group of collateral species. Such a schedule is in fact a faceted classification.

FACET ANALYSIS

The technique by which the terms subsidiary to a particular genus are divided into groups is called by Ranganathan "facet analysis". Its principles may be briefly illustrated.

In essence it consists of taking each of the subsumed terms and defining it, *per genus et differentiam*, with respect to its parent class. Thus, let us consider five terms from the field of chemistry: alcohol, liquid, volatility, combustion, and analysis. The first is a *kind* of chemical substance the second a *state* of that substance, the third a *property* and the fourth a *reaction* of it, and the fifth an *operation* performed by man upon it.

The five italicized words are the characteristics of division by which the terms are derived from the class chemistry. Looked at in another way, these characteristics are the logical categories by which the various terms are assembled. The examination of other fields of knowledge reveals other categories: *organ* in biology, the *use* of a crop in agriculture, the *agent* causing a disease, the *time* and *place* of an historical event and so on.

There has been some discussion in the Classification Research Group as to whether a limit can be set to the number of categories needed to cover the whole field of knowledge. In the Colon Classification, all facets are regarded as manifestations of one of five "fundamental" categories, named Personality, Matter, Energy, Space and Time. This point of view has not been adopted by the Group. We have not assumed that there is necessarily only a small number of basic categories, which figure in every field of knowledge. Nevertheless, we recognize that the use of a provisional set of categories - formulated by a practical examination of certain subject fields - is helpful in making a first approach to other subjects. As knowledge grows, the set of categories may need modification.

It should be mentioned here that another technique of analysis, developed by Farradane, has also been found useful. In essence it consists in the postulation of a set of basic "operators", representing possible relations between concepts. A series of compound subject headings is analysed by this means, the appropriate operators being inserted between successive terms in each heading. When the set is examined as a whole, there are seen to be groups of terms which have similar "relational" contexts, and these groups in effect constitute categories similar to those constituted by facet analysis.

NOTATION

A second criticism made of the classified index is that its notation may be "complex, confusing, lengthy and hard to remember". The Classification Research Group is very much aware of this problem. Although taking the view that the formulation of detailed principles of classification structure must precede close considerations of notation, the Group has undertaken some preliminary exploration of the latter field, and the following points have emerged.

(a) *Function.* The functions of notation are, in order of complexity:

- (1) in a random code scheme, to serve as a location reference;
- (2) in a systematic classification, to serve as a location reference, and to mechanize the ordering of classified material;
- (3) in a systematic classification, to serve as a location reference, to mechanize the ordering of classified material, and to represent certain features of the classification structure. The structural features which can thus be represented are:
 - (a) class hierarchical relations between terms;
 - (b) various classificatory dimensions of a subject, including the relations of one subject with another.

The notation of the Library of Congress Classification, and of the Bibliographic Classification are primarily examples of case (2) above; that of the Decimal Classification is mainly an example of case (3a); while the notation of the Colon Classification represents cases (3a) and (3b).

(b) *Criteria.* The criteria required in a notation are brevity and the use of symbols between which ordinal relations are self-evident. The Group has noted that in the Cordonnier and de Grolier syllabic notations, pronunciability rather than brevity has been taken as a primary requirement.

(c) *Mechanics.* The basis of notation must be one or more series of numerical or literal digits, the members of which are by custom arranged in an accepted order. The use of a single series will satisfy the requirement as to self ordinal relations more fully than the use of a mixed system or more than one series, but a mixed system has greater potentiality for brevity than the single series system. Arbitrary symbols of assigned ordinal value add to the complexity of a system, but a long notation which includes them may be more intelligible and easily carried in the memory than one of the same length consisting only of numerals or letters or both.

A notation should include mechanisms permitting the insertion of a new term at any point of array or in chain, or of a new series of terms representing a new dimension or facet of the subject. No existing system attempts this fully.

Systems of type (3) above express subordinate hierarchical relationship by adding digits to express division. Thus, the sign DE represents a topic which is species to the genus represented by D. Interpolation can be achieved only at the cost of some special modification of this principle, such as the Colon octave device (by which digits 8 and 91 represent co-ordinate topics), or the addition of a digit of a new species. In the type (2) systems the same devices may be used but do not represent a modification of any earlier principle adopted.

Systems of type (3b) may combine elementary terms representing the components of composite subjects, or they may combine terms to represent two or more subjects in relationship. The point of junction between combined symbols can be indicated by commencing the second term with a distinctive connecting symbol.

CONCLUSION

As has been stated earlier, this paper is intended only to give a broad picture of the views of the Classification Research Group, and to present an interim report on certain topics under discussion. The Group, after a long period of theoretical discussion, is now undertaking a detailed practical examination of a series of subject fields. In doing this, we will have the double purpose of further developing principles and of constructing satisfactory schedules.

The Classification Research Group is convinced that many weaknesses of existing classifications and other methods of information retrieval are due to an insufficiently searching consideration of principles on the part of their makers. Other weaknesses are due to the failure of an actual classification to conform with its stated principles.

We believe that a study of the foundations of subject analysis, such as this and other groups are now carrying out, will make our catalogues, our indexes, our classifications and our selectors more efficient tools in reference service. We by no means wish to work in isolation from others, and therefore welcome the suggestion that Unesco might support the development of a standard classification. We hope our own contributions will help forward that development.

MEMBERS OF CLASSIFICATION RESEARCH GROUP, 1ST JANUARY, 1955

Mr. Booth, T.I.D.U., D.S.I.R.
Dr. D.J. Campbell, Chester Beatty Research Institute.
Mr. E. J. Coates, British National Bibliography
Mr. J. E. L. Farradane, Tate & Lyle Ltd.
Mr. D. J. Foskett, Metal Box Co., Ltd.
Mr. G. Jones, Birmingham School of Librarianship.
Mr. J. Mills, North-Western Polytechnic School of Librarianship.
Mr. T. S. Morgan, Brixton School of Building.
Mr. B. I. Palmer, Education Officer, Library Association (Chairman).
Mr. O. W. Pendleton, Chartered Insurance Institute.
Mr. L. G. M. Roberts, Gas Council.
Mr. B. C. Vickery, Imperial Chemical Industries, Ltd., (Secretary).
Dr. A. J. Walford, Ministry of Defence.
Miss K. E. Watkins, Patent Office.
Mr. A. J. Wells, British National Bibliography.

This paper was first published in the *Proceedings of the International Study Conference on Classification for Information Retrieval*. London. ASLIB, 1957. It has been reproduced here with permission for ASLIB, the Association for Information Management, Information House, 20-24 Old Street, London EC1V 9AP, United Kingdom.

CLASSIFICATION IN INFORMATION RETRIEVAL: THE TWENTY YEARS FOLLOWING DORKING

E.J. Coates

Formerly Editor, British Technology Index, London
Present address: 11 Mount Pleasant,
St Albans, Herts. AL3 4QH, United Kingdom

HERBERT COBLANS AND CLASSIFICATION

MY CONTACT with Herbert Coblans was limited, alas, to our shared concerns in the field of subject indication. He first sought me out in the early 1960s, ostensibly to find out about the indexing methodology of British Technology Index, then with its birth agonies not too far behind it, but more probably to give moral support. From such a quarter this meant a great deal. I had, for some reason, expected to find him a mechanizing man, to whom I would be required to hand over copious rudimentary enlightenment on subject indication questions, but this was quite wrong. It was I who soon found myself on the receiving end of the enlightenment process.

It appears from the bibliography of his published writings that he did not, at any time, write a paper wholly devoted to classification in any of its aspects. However, in his many contributions to the literature of complete information systems, subject indication, including classification, is invariably treated as an element of significance. In this he stood outside a fashion, prevalent in the early flush of mechanization, of placing subject indication among the low-priority requirements of complete systems. As a linguist of considerable accomplishment, and with his extensive experience in helping to push ideas across language barriers, he had a firmly grounded sense of the autonomy of concepts aside from the words or other linguistic units employed to represent them. It was perhaps for this reason that he took the view that concept organization, a human activity, was an essential factor in an information system, whatever automatic indexing might eventually have in store.

One view of classification research held within the information professions is that it has much of the character of a game played by a few devotees for its own sake, rather than as a practical response to an identified need. Coblans may well have agreed with this view, up to a point. If so, he would have at the same time have taken the position that there was a possible spin-off from this particular game which would be, in a practical sense, relevant and non-trivial. His own relation to the game was rather that of an extremely well-informed spectator, whose grasp of the game's finer points was such that he was often well-placed to supply valued counsel to the players. He was frequently in demand to preside over committees and larger gatherings which had some aspect of subject indication or classification as their focus. As a discussion topic classification often generate unruly emotions which can be turned to constructive account only by interaction with authority of the kind which Coblans exercised. His contribution to classification in this role alone was considerable.

A quality which he particularly brought to bear upon the subject was a more than average awareness of the temporal aspects of things, including an awareness of history. He was, for instance, no stranger to the contradiction at the heart of the management problems of classifications and subject indication languages; namely, that as tools for the preservation of information availability over time, they necessarily freeze the flow of time as snapshots of a state of knowledge at a particular moment, yet on the other hand they need a ready capacity to be dynamic, to change at minimal cost and inconvenience. As to his sense of history, this it was which enabled him to recognize what was worthwhile in the conventional and established bibliographical classifications, while at the same time welcoming new classifications which would profit,

with appropriate hindsight, from the revealed shortcomings of their predecessors. He spoke to me once, not very long after his retirement from full-time work, of the swiftness with which prominent professional names entered oblivion after their owners had departed from the scene. If, as I think he did, he included himself, then perhaps it is permissible to think that for once his modesty may have caused him to be mistaken.

CLASSIFICATION SINCE THE DORKING CONFERENCE: THEORY AND IMPLEMENTATION

The International Study Conference on Classification for Information Retrieval held at Dorking in 1957¹, is a convenient take-off point for an attempt to discover if there has been, over the last two decades, any coherent trend in the direction of classification activity and in the manner in which it has been viewed by the information and library professions.

Comparing this period with that which preceded it, two striking features immediately come to view. The first is that since 1957 there has been no further breakthrough in understanding of classification at a fundamental level comparable to that achieved earlier by Ranganathan. Secondly, there has been a very great expansion of practical activity in the making of new classifications, the updating of old ones, and in the production of indexing tools which include classificatory elements, either overtly or in the background.

Ranganathan's contribution may be summarized under two heads, using terms which are not his own, but which reflect thinking which received its initial impetus from him. First, he demonstrated the unity of subject indication as a concept, embracing classification and subject indexing of all kinds. Second, he uncovered for the first time (though this is not to say that others who preceded him had not made fragmentary sightings) the whole syntactic axis of subject indication, to which he attached the label 'facet'. To get this very substantial contribution into proper perspective, it is necessary to point out that in respect of the other, and perhaps even more fundamental axis of subject indication - the semantic axis - Ranganathan was a traditionalist who made no new contribution, apart from those fresh points of view which could be inferred from the very notion of the unity of subject indication.

In the period since 1957 such progress as has taken place on the theoretical side of classification has been exclusively in relation to the syntactic axis. On the semantic axis, viewed broadly, there has been a near-standstill, though possibly the notion of integrative levels may have stimulated re-thinking on the respective roles of subject disciplines and phenomena as concepts included in classifications. Perhaps also the unambiguous identification of the semantic and syntactic axes themselves, by Gardin in connection with the SYNTOL system,^{2,3,4} should be accounted as progress on the semantic side.

CLASSIFICATION SYNTACTICS

On the side of the syntactic axis, the early part of our period saw a steady process of refinement and elaboration of the original facet analysis of Ranganathan. Some of this refinement was the work of Ranganathan himself; others, such as Vickery,⁵ extended the facet categories in ways suggested initially by the literature of science and technology, but probably also applicable with little change to other areas of knowledge. But perhaps the most important step in the syntactic refinement process has been the realization that facets are themselves essentially the functions or superficial manifestations of relations between concepts belonging to different facet categories. It was observable that a facet structure defined exclusively in terms of categories of concepts (Energy, Matter, Place, etc. and however extended) did not always ensure unambiguity of meaning. It was also seen that where ambiguity was successfully avoided it was often only because of the context restriction of a bounded field. This led Gardin⁶ to conclude that universal facets transcending discipline boundaries in a general classification were essentially different from the facets of special classifications. Ranganathan's own Rounds and Levels were seen clearly to be dependent in the last resort upon relational considerations rather than upon those arising from division into categories.

The place of Farradane's^{7,8,9} relational indexing system, first propounded in 1950, within the stream of classification theory was at first unclear, both to its author and to others. By 1957, however, it was beginning to be realized that the relationally-derived 'analets' of Farradane approximated to the index structures represented by synthesized notation in faceted classification. A year later, Kergévant¹⁰ proposed a scheme for making explicit all the various relations comprehended by the relation symbol (colón) in the Universal Decimal Classification (UDC). In the first half of the 1960s there was much attention given to the possibility of using relational techniques as precision devices in post-coordinate indexing, under the rubric of 'roles'.¹¹ Thus was understandable as an attempt to recover information which post-coordination methodology otherwise sacrificed in eliminating syntactic word order. The use of 'roles' however was shortlived, partly, one suspects, because the 'roles' were somewhat uncertain mixtures of generic context indicators and true relational terms, and partly because links could often fulfil the same function. Relational indexing in conjunction with mechanization was envisaged in the SYNTOL model already mentioned. Later in the 1960s, Perreault¹² devised a scheme of relators for UDC more ambitious than that of Kergévant.

Of the proposals mentioned above, comprising or including syntactic relational indexing systems, the Farradane system and SYNTOL have sparse relator enumerations, firmly anchored to the idea or logical plane. On the other hand the lists of Kergévant and Perreault are more extensive, include category terms, and at times appear to be pitched on the verbal plane. SYNTOL may be contrasted with the Farradane scheme on the ground that its underlying ideas are linguistically oriented, whereas Farradane's ideas are derived from a psychological theory. Possibly this difference in origin heightens the significance of the resemblances between the two systems, as detected by De Grolier.¹³ It is perhaps worth making the point that a system may be linguistically oriented in its ideas and terminology (as indeed this paper is in using the expressions 'syntactic axis' and 'semantic axis') and at the same time concerned with concepts rather than language. The SYNTOL provision for syntactic indexing was part of a total indexing system at the outset. Farradane gave attention to the semantic axis as an adjunct to his relational system only in the 1960s,¹⁴ but did not elaborate it concretely. Neither his postulated hierarchies nor the 'réseau notionelle' of SYNTOL added anything to existing ideas on the semantic axis as such.

Nowhere is the primacy accorded during the period to the syntactic axis more forcibly illustrated than in the story of the (British) Classification Research Group studies on a new general classification from 1963 to 1968.¹⁵ This activity was supported by NATO - a circumstance open which Coblans' comments would perhaps have been particularly illuminating. The study produced no general classification, not even an embryo one, but did useful work further elaborating *general* facet categories, i.e. categories not encompassed as Ranganathan's had been, by individual disciplines. Relations were also considered, but rather as complementary descriptions of the categories, than as explanations of the categories at a deeper level. Some limited progress was also made in identifying the main types of material expected to occupy each category, and in finding an ordering principle based upon integrative levels. Austin¹⁶ suggests that the absence of a classification as an end-product was in some sense an effect of mechanization potentialities, but his list of six desiderata of a system for exploiting machine-held files does not seem to amount to more than a faceted classification with explicit relators, plus regrouping-on-demand facilities, presumably dependent upon some form of structure-expressive coding. Exhaustion of funds finally terminated the CRG general classification study at a stage at which 'neither the order in which these categories are set down, nor the contents of the categories themselves can be regarded as better than speculative'. However the results of the study were picked up and became the starting point of the development of a syntactic subject indexing system, namely PRECIS.

IMPLEMENTATION

The advances, at the research level, in classification syntactics have been on a modest scale since 1957. On the other hand, during the same period there has been a very considerable amount of activity in which practice has in effect been engaged in catching up with the theory laid down in the quarter century before 1957. The two strands in Ranganathan's thinking, subject indication as a unified notion, and the facets, to some extent remain separate in the realm of applications. Unified subject indication leads to innovations of a semantic character. Facets in theory lead to facets and syntactic indexing in practice.

In its time aspect, this catching up process perhaps illustrates the well known phenomenon that a lag of about thirty years occurs between a significant discovery or an important new idea and the general appreciation of its importance. The first edition of the Colon Classification appeared in 1933. By 1957 the Dorking Conference declared that faceted classification was *the* basis of all information retrieval (in retrospect the exclusive definite article appears unfortunate). By the early 1960s, several special subject classifications all embodying the facet principle had appeared, and in most cases were soon in active application.

By the present date, the catching up process - involving concrete systems capable of demonstrating the theory - has probably gone far enough for some residual problems, the starting points for further research effort to be identified, but not yet far enough to change the image of classification held in the information professions at large, which has been generated by long exposure to the flaws in the two chiefly established general classifications for library purposes, the Dewey Decimal Classification (DDC) and the Library of Congress Classification.

The overtaking by practice of the theory of the earlier period has become visible in manifestations of two kinds. In the first kind the product is a classification system. In manifestations of the second kind, the primary product is an alphabetically ordered indexing language, descriptor list, thesaurus, controlled vocabulary, or a system for producing alphabetically ordered pre-coordinated syntactic strings. Associated with these primary products are classificatory parts or ancillaries: these are most commonly classificatory fragments, sometimes rudimentary classification schemes, and in a few cases fully-developed classification schemes.

With regard to the classification schemes in their own right which have appeared in the period under review, the early years of the period were, as has been mentioned, prolific in special subject classifications, too numerous for individual mention. Some of them remain operational.

IMPLEMENTATION: GENERAL CLASSIFICATIONS AND A GENERAL SUBJECT INDEXING SYSTEM

It is also necessary to take into account both new general classification schemes and the extent to which the new (but pre-1957) ideas have permeated into successive revisions of the established schemes. In this area there has been one undoubted success and one evident disaster.

It is not too early to speak of the second edition of the Bliss Bibliographical Classification (BC2)^{17,18} as a major success, even though so far only four of a projected twenty parts have been published. The parts which have appeared (Introduction and Auxiliary Schedules, Education, Social Welfare, Religion) are a success on their own account. Here, perhaps for the first time, can be seen, at the detailed level, the embodiment of most of the theory of classification from 1933 onwards. Revision is an inappropriate term to give to the process of inserting a fully faceted structure into the remarkably accommodating shell bequeathed by Henry Bliss. BC2 combines the syntactic versatility of the Colon Classification, with much of the BCI philosophy on simplicity of notation. It is, at the moment, up to date in content, and in detail pitched somewhere between monograph and journal article level, often nearer the latter. Reception of the parts so far available suggests that despite the claims of the established classifications as *de facto* standards, and despite the view that all classifications are defective and that there is little to choose between them, there is also a decided hunger for the best that modern classification can produce. When the decision to revise BC was taken, it was questioned whether any library would adopt the revised scheme, whatever its merits. Now that the first parts have appeared, it seems likely that BC2 will not lack for takers. In broadest terms, the real achievement of BC2 is that it has gone so far in closing the gap between classification theory and practice, that it demonstrates the theory in real life, and thus permits assessment of such factors as the cost of applying it, its communicability for teaching purposes, and the satisfaction offered to (and accepted by) the ultimate user. The second of these factors could be of greater than usual importance in the present case. A generation of students in library and information science have been taught classification theory, only to find that there is no general classification scheme which fully shows it forth. BC2 is in process of changing this, but as will appear from what follows, only just in time.

Up to the end of 1970, the nearest approach to a model of the working of faceted classification in the total field of knowledge was that offered in the current and cumulative issues of the British National Bibliography (BNB). BNB is a classified bibliography, primarily of British monograph publications, started in 1950. The scheme of classification was DDC 14th edition, supplemented as time went on by selective adoptions from the 15th, 16th, and 17th editions. The input includes a substantial element of monographic report material and the degree of specificity required was therefore often documentation level. No DDC edition has given detail corresponding to specificity at book level, let alone the level required to characterize BNB input. For its own purposes BNB expanded DDC to the level required, by the methods of facet analysis. Furthermore, in the frequent instances where DDC's policy was one of permissiveness through reticence or silence, BNB chose a direction governed by considerations of faceting and consistent facet order. At first this superimposition of a facet structure upon a mainly non-faceted, but sufficiently indeterminate, scheme was carried out without additions to the DDC notation, but later a supplementary notation was added to meet the requirement that the alphabetical subject index should indicate specific addresses. Thus, for twenty years BNB carried the banner - albeit a banner with a message subject to distortion by the constraints of DDC - for faceted classification. It was undoubtedly the main channel by which Ranganathan's ideas became familiar to the world outside India, and its influence on the practice of classification everywhere apart from North America was incalculable. The fact that the word 'facet' has become an international word used by documentalists in many different languages has been due, more than to any other factor, to the twenty-year demonstration of facets in practice by BNB.

The disaster to the cause of classification was that this demonstration was discontinued after 1970. One has only to compare the generally ordered and well knit structure of the BNB for any year in the 1960s, with the mediocre, weakly structured arrangement, with frequent and unaccountable back-looping, that is now offered, to appreciate that in terms of what Perreault¹⁹ aptly terms 'transparency' the change is, without overstatement, a disaster. Opacity has become the norm. For someone pursuing a chronological search in forward direction, starting in the 1960s and passing over to the 1970s, making use of the classified section of BNB to survey and extract bibliographical information on a moderately broad subject area, and hoping to pick up further useful material marginal to the target subject, the going becomes distinctly harder after passing the 1970/71 threshold. A study to quantify this particular transparency/opacity transition in terms of search time would be feasible and useful.

The post-1970 BNB comprises a classified section using DDC 18th edition unsupplemented, with verbal features derived from PRECIS strings, plus a PRECIS subject index. In effect the use of the ordering of the classified sequence as a search aid is discontinued, and the PRECIS index, independent of the classification, becomes the main means of retrieval. The classified section serves in relation to searching only as an addressed store of full bibliographical descriptions of items. Thus the role of the classification as a search pathway has been dropped, or at least seriously degraded, while the subject index which formerly used chain procedure has been enhanced by a PRECIS system, which produces, by a form of cyclic sequencing of terms essentially in the manner of KWIC, fully specific entries at every alphabetical access point. There is no question about this enhancement. It is a considerable practical asset in a manually scanned system, if one can ascertain at a single step that the system store does *not* contain the target material, and this is what PRECIS permits. On the other hand, the facility for broad area survey or search, and for browsing, by patterned collocation - the strong point of the pre-1971 BNB, and indeed the strong point for classifications generally over alphabetical indexing languages - has been to all intents and purposes abandoned. It is true that DDC 18th edition moved to some extent in the direction of faceting, but not nearly far enough to compensate for the loss of BNB's own faceted extensions. This classification just does not have the integrity of structure needed to sustain the needs of subject retrieval in a national bibliography.

It was stated earlier that one type of manifestation of the passing of classification theory into practical application consisted of classificatory fragments associated with alphabetically ordered indexing languages. PRECIS²⁰ is not an indexing language in the usual sense which assumes a predetermined and authorized list of terms. Essentially it is a set of procedures for producing an alphabetical subject index in which the entries, when they are not simply single words or phrases, consist of syntactic strings. It claims to be free of all classifications, but the strings may legitimately be considered as classificatory syntax. Formally it is a relational system, though some of its 'roles' are defined in categorical terms. 'Roles' are determined at

the concept level, and semantic control of the terms linked by the 'roles' is aided by reference to an associated semantic cross-reference file comprising synonym-couples and classificatory semantic-couples of related terms. Choice of hierarchies at levels where there are many alternative possibilities is apparently made on an intuitive basis. From a single input string, the computer outputs syntactically ordered strings, for as many significant terms as there are in the string, each of such terms occupying a leading position.

PRECIS is the most elaborate, but not the only relation-based syntactic indexing system developed for operational use during the 1960s and computer processed. Two systems of the same general kind, though producing very different output, which antedated PRECIS were that employed by British Technology Index (BTI)²¹ from 1962 onwards, and the Articulated Index system of Armitage and Lynch,²² later used in Chemical Abstracts services. The BTI system was computerized for editorial procedures, including manipulation of syntactic strings and extraction of semantic cross-references, in 1968. The input to the Armitage and Lynch system is a title-like phrase and a primary entry word. The computer then uses prepositions as cues for forming intelligible permutations needed as index entries. This system is of interest as the only one in which relation identification process is carried out at the verbal level, though it could be argued that the formulation of the title-like phrase involves a degree of concept analysis.

PRECIS has been reviewed at some length in this journal,²³ and no detailed account need be given here. From the point of view of illuminating general problems in classification, its residual quirks and occasional subterfuges are of interest. What are called 'Differencing operators' represent an indexing situation, which though recognized before, for instance in SYNTOL, is dealt with more fully in PRECIS than elsewhere. These operators are used in the handling of many kinds of adjective+noun or noun+noun phrases. In such cases the two words in conjunction represent the species of which one of the words represents the genus (e.g. in the phrase CHAIN SAWS, SAWS is the genus of CHAIN SAWS). We have here a situation which well illustrates the vital need to distinguish linguistic formal relations from logical relations. In the literature of indexing the genus/species relation is normally associated with the semantic axis, but in instances of the present kind this same genus/species relation appears on the syntactic axis. Classification is concerned primarily with the logical relation, and it is immaterial whether the species concept is expressed in language by a single word or by a syntactic combination. However, from the point of view of indexing (including PRECIS and any alphabetical subject index to a classified file or classification schedule) questions of linguistic form may be critical. What emerges from the manner in which PRECIS handles the 'difference relation' (e.g. the relation between SAWS and CHAIN in the input construction (1) SAWS (\$i) CHAIN) is that in principle *any* syntactic string could be interpreted as a system of 'difference relations', as soon as one begins to inspect, not the relations between the string elements A, B and C, but those between A and B+A, or between C and A+B+C. The 'Difference operator' is clearly a necessity, but it could be a ready source of ambiguity if its use is not severely circumscribed. The argument used to justify its extended employment for handling all the prepositional phrases that do not fall into the territory of operators (1) to (3) does not seem to stand up to examination and suggests that the scheme of 'roles' needs elaboration. Of all of Ranganathan's categories 'personality' has proved to be the most elusive. Possibly further study of the PRECIS 'difference relation' could offer a clue as to what it is.

In the sphere of general classification some signs of the trend towards syntactic sophistication can be seen in the revision policies both of DDC and UDC, and in the recent arrival of one further new classification, the Broad System of Ordering (BSO).²⁴

Sufficient indication has already been given, in connection with the discussion on the 1970/1971 threshold in BNB's practice, that the programme of modernizing DDC, welcome though its general direction is, has not gone very far. Nevertheless, some features of BNB's pre-1971 practice, of which DDC is said to have disapproved at the time, have turned up in the 17th and 18th editions of DDC in the authorized and approved livery. The UDC, which practiced notational synthesis long before the emergence of faceted classification and the syntactic axis in classification, has gone further in catching up with the theory of 1933 to 1957, and perhaps it was for this reason that in a review by an Aslib research team,²⁵ undertaken as a run-up to the Unisist Study Report,²⁶ on the suitability of existing classifications for the subject indication switching role within the Unisist network, UDC was singled out as the least defective system. The continuing interest on the part of UDC circles in relational analysis has already been mentioned, and there were signs in the late 1960s of the emergence of a fully conscious and coherent policy trend in the

direction of more faceted schedules.²⁷ That the policy has, to date, been implemented less fully than might have been hoped, has been due less to problems arising from the policy itself than to difficulties ostensibly of an administrative character. However, these difficulties probably have their roots in the lack of a sufficiently coherent general theory of classification which would command a reasonable consensus. Detailed decision making which depends upon laboriously achieved collective intuition rather than or delegated application of generally agreed and understood principles, can keep pace with knowledge only with the greatest difficulty.

Though UDC was not given the task of acting as a switching language to UNISIST, its parent body, the International Federation of Documentation (FID) was given an assignment, under a series of contracts with UNESCO, to compile a new broad classification or system of ordering for the switching purpose. The networking concept itself was triggered early in the 1960s, by the prospect of technological advance which would lead to cheap and efficient data transmission facilities. An underlying assumption was that the individual centres within a network would not, and should not be expected to, adopt a common indexing language which they would all apply to documents. Local indexing languages in use were often optimal for particular local purposes. Transmission of subject information from one centre to another would therefore call for the use of a communication or switching indexing language which would mediate between the local languages. Compatibility by means of direct concordance tables was appropriate if there were two languages only in use in the network. If there were three languages direct concordance or concordance via a neutral switching language were equal in terms of the number of concordance tables required. If there were more than three local languages used in a network, then connection through a switching language was economically advantageous. Studies by the Groupe d'Etude sur l'Information Scientifique at Marseilles^{28,29} had suggested that such a communication language, which should be structured, might operate without too much transmission loss, and quantitative confirmation was later obtained as a result of a research project (of the Advisory Committee of which Coblans was the first Chairman) at the Polytechnic of North London School of Librarianship.^{30,31}

The Broad System of Ordering was conceived both as a structured mediating language which might operate in a context such as is outlined above, and as a subject code for use in conventional disseminative media, such as the publications of specialist institutions and directories of such institutions affording subject access or indication. The needs of developing countries were to be kept in mind, the scheme was to be amenable to mechanization, and as far as possible such aspects as the effort of management and updating and the effort of using such a scheme at an individual centre were to be taken into account. In a paper which forms part of a tribute to Coblans it is pertinent to add that the BSO concept had his support, and that at his request he was kept posted informally on its progress.

BSO, like BC2, incorporates many of the advances in the theory of the last half century. It is faceted, with provision for facet combination across discipline boundaries under the control of relational considerations. It is based upon disciplines but also provides for entities and phenomena treated in a multidisciplinary manner. It was compiled on the basis of institutional warrant, not documentary warrant, the cut-off principle employed being that a subject which has an organized information source devoted to it solely should be represented specifically in BSO. This principle led to a compilation of just over 4000 terms. The design attempts to optimize on the factor of ease of updating and maximum predict ability in dealing with future new knowledge. The 1978 BSO has a non-expressive numeral notation devised for human users. A completely hospitable, completely expressive notation, repulsive to humans, but possibly advantageous for computer handling, could be substituted for the present notation without disrupting the structure of the system.

CLASSIFICATION SEMANTICS: APPEARANCES IN PRACTICE

There has been no question of practice overtaking theory on the semantic axis during the period under review. There has been no theoretical advance to be overtaken. However, the period has not been lacking in fresh manifestations of a semantic kind which seem to vindicate Ranganathan's view on the unity of subject indication, that alphabetical subject indexing schemes possess, or should possess, some form of classificatory correlate. At the beginning of the period it seemed that the classificatory correlate was going

to be disproved in practice. Pioneer mechanizers of the preceding period, contemplating established manual subject indication systems, could find no programmable logic in them, and concluded, correctly, that they must be intuitively based. Accordingly, the earliest mechanically aided systems minimized the human process of vocabulary control or eliminated it altogether. The first post-coordinate systems depended upon the simplest possible initial assignment of keywords; thereafter the whole process was mechanically aided clerical work. However it was soon seen to be necessary to add such humanly selected refinements as a generic posting. This typifies a trend which was to be repeated on a grander scale when computers took over from mechanical and optical coincidence aids. Initially vocabulary control processes, including classification, are minimized. Then in a second phase they return, more or less openly, perhaps as an adjunct to the computerized part of a system. These adjuncts, in the shape of indexing languages, are rather more commonly placed on the man side of the man/machine interfaces than processed within the computer. Hence the increasing emphasis on the use of thesauri as search aids, especially in interactive dialogue.

The whole period has seen an unprecedented expansion in the making of thesauri, intended mainly for post-coordinate systems. The essential differences between thesauri and their predecessors the subject heading lists lie first in the relative paucity of compound headings admitted in thesauri, and second in the more explicitly classificatory form (BT) and (NT) in which thesaural crossreferences are given. These cross-references are in effect classificatory fragments scattered by the alphabet. On trying to piece these fragments together, one does not usually arrive at complete classification systems, even if (RT) cross-references are included as possible collateral links of a classification. From the middle of our period there has been an increasing tendency on the part of thesaurus makers to include schedules of subject groupings of various kinds to supplement the main alphabetical sequences of descriptors. These groupings range from broad categorizations, often apparently prepared by manual means, tree diagrams, full scale alphabetic-classed configurations prepared by computer manipulation, through to full-scale classifications. Thus in the wake of mechanization classification appears to be seeping back into the picture. No conscious attempt is being made in these cases to apply a theory of the classificatory correlate to an alphabetical subject index, but it has been found that displays of terms physically collocated by meaning form useful supplementary aids to searchers, not to mention vocabulary controllers and index language updaters. Van Slype³² has 'Classification of the descriptors' as first item in a table of thesaural characteristics.

The cross-reference systems of modern thesauri present two problems which point urgently to the need for further research in the neglected area of the semantics of classification. The first concerns thesaural (RT) references which standardization activities nor the work of terminologists have satisfactorily elucidated. The second relates to the apparently arbitrary incidence of polyhierarchies on the (BT) (NT) axis. On the one hand monohierarchy may well be a necessary property of classifications utilizable as linear displays of concept relatedness on two-dimensional media. On the other hand, the additional hierarchies may be shown to be useful search pathways. If may be asked why the (RT) and polyhierarchical references should not remain elusive. The short answer is that the control of the vocabulary itself will to that extent be partial. Again it may be argued that such loopholes in control do not matter greatly in day-to-day use of the language. The point may be conceded in relation to day-to-day indexing and retrieval, but not in relation to operations upon the indexing language itself, such as updating and transformation (e.g. translating the terms of a thesaurus into those of a different natural language, linking with a switching language). Such operations are likely to be of increasing importance.

CLASSIFICATION UNDER FIRE

Suggestions that classification for information retrieval is obsolete or of dubious utility have come from two quarters - that associated with mechanization, and that associated with system evaluation.

It was at first suggested that mechanization as such would make classification unnecessary. As has been shown earlier, classification, though not classification systems, made its comeback often in the context of mechanized systems themselves. Another line of criticism associated with mechanization has been that classifications cannot profitably be brought into mechanized systems because they do not make the fullest

use of computer capabilities. This complaint is sometimes backed by a call for a new radically non-traditional form of classification. Such demands can be interpreted as demands for classifications with a greater number of systematic features than present classification systems possess. There has been progress in classificatory syntactics, and as a consequence many systems predominantly syntactic in nature have been mechanized. On the other hand progress in the systematization of classificatory semantics has been almost zero. Accordingly there is not very much for computers to do in connection with the semantic side of classifications as we now know them. Urgent priorities for research on thesaural cross-references have already been mentioned. They are relevant to this complaint.

The case against classification on the basis of evaluation tests is a corollary of the conclusion from such tests that controlled indexing languages perform hardly, if at all, better than uncontrolled indexing. Classifications are forms of controlled indexing language. Therefore this same conclusion goes for classifications. Furthermore comparative tests including classification systems have substantiated the general finding. Keen,³³ who puts forward views on these lines, agrees, however, that all the factors involved in the use of classification for information retrieval have not been covered in tests, and adds that, for example, the systematic arrangement of compound and complex topics in the form of a helpful sequence has not been tested. Such a proviso rather removes the sting from the evaluators' scepticism. Perhaps the last word on the central place of helpful sequence in all considerations as to the utility of classification may be left to Perreault³⁴ '... if we have not found what we need, what do we do next?'

REFERENCES

1. INTERNATIONAL STUDY CONFERENCE ON CLASSIFICATION RESEARCH. Dorking, 1957. *Proceedings*. London: Aslib, 1958.
2. Gardin, J.-C. and Levy, F. Le Syntol (Syntagmatic Organisation Language). In: *Information processing 1962*. Amsterdam: North Holland Publishing Co., 1963, pp. 279-83.
3. Gardin, J.-C. *SYNTOL*. Rutgers State University, Graduate School of Library Service, 1965 (Rutgers series on systems for the intellectual organization of information, edited by S. Artandi. Vol. 4).
4. Cros, R.C., Gardin, J.-C. and Levy, F. *L'automatisation des recherches documentaires. Un modèle général: le SYNTOL*. 2nd edn. Paris: Gauthier Villars, 1968.
5. Vickery, B. C. *Faceted classification*. London: Aslib, 1960.
6. Gardin, J.-C. Free classifications and faceted classifications: their exploitation with computers. In: *Classification research*. Proceedings of the Second International Study Conference, Elsinore, 1964. Copenhagen: Munksgaard, 1965, pp. 161-8.
7. Farradane, J.E.L. A scientific theory of classification and indexing: further considerations. In: *Journal of Documentation*, 6, 1950, 83-99.
8. Farradane, J.E.L. A scientific theory of classification and indexing: further considerations. *Journal of Documentation*, 8, 1952, 73-92.
9. Farradane, J.E.L. Analysis and organisation of knowledge for retrieval. *Aslib Proceedings*, 22, 1970, 607-16.
10. Kergévant, D. Développement de l'analyse des relations dans la CDU. *Quarterly Bulletin of the IALD*, 3, 1958, 111-16.
11. Costello, J.C. A basic theory of roles as syntactical control devices in coordinate indexes. *Journal of Chemical Documentation*, 4, 1964, 116-24.
12. Perreault, J.M. *Towards a theory for UDC*. London: Bingley, 1969, pp. 133-48.
13. Grolier, E. de. *A study of general categories applicable to classification and coding in documentation*. Paris: Unesco, 1960.
14. Farradane, J.E.L. Concept organisation for information retrieval. *Information Storage and Retrieval*, 3, 1967, 297-314.
15. Classification Research Group. *Classification and information control*. London: Library Association, 1970 (Library Association Research Pamphlet No. 1).
16. Austin, D. Trends towards a compatible general system. In: Maltby, A. (ed.). *Classification in the 1970s*. London: Bingley, 1972, pp. 211-48.

17. Bliss Bibliographic Classification: 2nd edn. edited by J. Mills and V. Broughton. London: Butterworths, 1977. in progress
18. Roberts, N. Review of Bliss Bibliographic Classification: 2nd edn. *Catalogue and Index*, 46, 1977, 5-7.
19. Perreault, J.M. *op.cit.* p. 90.
20. Austin, D. PRECIS: a manual of concept analysis and subject indexing. London: Council of the British National Bibliography, 1974.
21. Coates, E.J. British Technology Index. In: *Encyclopedia of library and information science*. New York: Dekker, 1970, Vol. 3, pp. 327-341.
22. Armitage, J.E. and Lynch, M.F. Some structural characteristics of Articulated Subject Indexes. *Information Storage and Retrieval*, 4, 1968, 101-11.
23. Coates, E.J. Review of Austin D., *Precis: a manual of concept analysis and subject indexing*. *Journal of Documentation*, 32, 1976, 85-96.
24. BSO: *Broad System of Ordering: schedule and index*. The Hague: Fédération Internationale de Documentation. Paris: Unesco, 1978.
25. Aslib. Research Dept. *Classification in science information: a comparative study undertaken for the International Council of Scientific Unions*, 1969 (Document UNISIST/CSI/5.8).
26. UNISIST: *study report on the feasibility of a world science information system*. Paris: Unesco, 1971.
27. Arntz, H. Die DK - eine Vielfacettenklassifikation. *Nachrichten für Dokumentation*, 21, 1970, 139-42
28. Gardin, N. and Levy, F. *Lexique documentaire pour l'information scientifique*. Paris: Groupe d'Etude sur l'Information Scientifique, 1967.
29. Gardin, N. The intermediate lexicon. *Unesco Bulletin for Libraries*, 23, 1969, 58-63.
30. Horsnell, V. *Intermediate lexicon for information science: a feasibility study*. London: Polytechnic of North London School of Librarianship, 1974.
31. Horsnell, V. and Merrett, A. *Intermediate lexicon research project: phase 2, evaluation of the switching and retrieval performance of the Intermediate Lexicon for Information Science*. London: Polytechnic of North London School of Librarianship, 1978.
32. Van Slype, G. Qualitative and quantitative characteristics of monolingual and multilingual thesauri. In: Third European Congress on Information Systems and Networks, 1977. *Proceedings*, Vol. 1. Munich: Verlag Dokumentation, 1977. pp. 403-15
33. Keen, E.M. Prospects for classification suggested by evaluation tests carried out 1957-1970. In: Maltby, A. (ed.). *Classification in the 1970s*. London: Bingley, 1972. pp. 193-210.
34. Perreault, J.M. *op.cit.* p.198.

COATES: This paper first appeared in *Journal of Documentation* Vol. 34 No. 4, December 1978 p. 288-299 (ISSN 0022 0418) and has been reproduced here with kind permission of the original publisher, ASLIB, the Association for Information Management, Information House, 20-24 Old Street, London EC1V 9AP, United Kingdom.

NEW TOOLS

The Dorking Conference produced a "manifesto" (reproduced above) under the title "The need for a faceted classification as the basis of all methods of information retrieval", which continued to be a rallying point for the CRG in its fight for funding but which was also clearly demonstrated in such practical artefacts as the second edition of the Bliss Classification, PRECIS, the Thesaurofacet and ISO 2788, the international standard on compilation of thesauri.

The two papers that follow review structure and function in retrieval languages and knowledge representation, while the third demonstrates the impact (and some of the problems) of natural language processing.

Alan Gilchrist

STRUCTURE AND FUNCTION IN RETRIEVAL LANGUAGES

B.C. Vickery

Aslib Research Department

ABSTRACT

Retrieval languages may have varied structural characteristics, and these are summarized. The languages serve varied purposes in information systems, and a number of these are identified. The relations between structure and function are discussed and suggestions made as to the most suitable structures needed for various purposes.

When the varieties within a group of entities are first distinguished - whether the entities be animals, rocks, chemical substances, political institutions, or whatever - it is always the most obvious and superficial characteristics that are to begin with chosen as a basis of division. Subtler and more vital characteristics are not recognized until later. This course of events has been true of the analysis of languages for subject retrieval.

For centuries there have existed indexes and classifications. The former used letters and the latter often used numerals, and this obvious difference in form long seemed significant. Classifications with letter or mixed notation confused the issue, but the purely notational characteristics of "decimal" and "colon" served to name two famous classifications. The fact of an underlying difference in structure, as between alphabetic and classified arrangement, was of course well understood.

Closer analysis of structure has developed as a consequence of deeper study of the functions of retrieval languages. Initially, two functions were recognized: to identify, and hence to select items on a specific topic (the task of the alphabetic index); to group in proximity items on similar subjects (the task of classification). The two functions (specific reference and generic survey) seem opposed. Ranganathan in particular, urged the provision of both by a helpful symbiosis between classified and alphabetic arrangement.

The structures of both types of language began to grow more complex. Classificatory structure was explored and developed by Ranganathan and other advocates of the faceted approach. Alphabetical terminologies became thesauri. With the coming of computer retrieval, new complexities and new simplicities appeared: on the one hand, highly structured codes like that of Western Reserve University; on the other, systems using natural language titles or abstracts.

What functions are served by these new structures? They are believed to be related to the effectiveness of retrieval, and thoughts have turned to measures of this effectiveness. Perry and his colleagues suggested several measures, such as recall and pertinency factors, which have later been put to good use by Cleverdon as recall and precision. Retrieval languages may be analysed into structural elements, whose function is related to effectiveness. Some "devices", when introduced into an uncontrolled vocabulary of simple terms, tend to increase recall, others to increase precision.

The structural elements identified at this stage include, as *recall* devices: confounding of true synonyms, of near synonyms, and of word forms, and use of more generic terms; and as *precision* devices: co-ordination of terms, interlocking of terms in a topic, expressing relations between terms in a topic, and weighting index terms. Hierarchical linkage of generic and specific terms aids both recall and precision.¹

A quantitative approach has also been developed: a simple measure is the number of separate terms in a retrieval language, but this has to be related to the scope of its subject field. Some ratio of terms to items in the field seems a more suitable measure of the average specificity of the terms. Other aspects can be quantified - for example, the average number of links in hierarchical chains, or the average number of cross-references in a thesaurus.

A linguistic analysis of retrieval languages by Coyaud² emphasizes their similarities (or lack of similarities) to natural language. The views of a computer programmer, Meadow, make an interesting contrast.³

All these approaches to the analysis of retrieval language are of continuing value. Some practical studies of computer information systems undertaken by Aslib Research Department have suggested a further approach, which is reported here.

STRUCTURAL CHARACTERISTICS

Let me first recapitulate some structural features of retrieval languages. It is helpful to distinguish between an *indexing* and a *search* language. In any particular information system, a common retrieval language underlies both, but each may also include elements special to itself. Secondly, the notion of a *dictionary* converting from one language to another needs to be given more emphasis. Thirdly, the notion of a *thesaurus* linking semantically related words within a particular language needs to be refined.

The building blocks of a retrieval language are *terms*: These may be words, codes, or other symbols. It is convenient to distinguish between 1) terms occurring in the texts to be indexed and 2) terms assigned to the texts. Some retrieval languages consist wholly of the first (e.g. natural language titles), some wholly of the second (e.g. class numbers), and some are a mixture (e.g. terms drawn from a thesaurus).

The multiplicity of terms occurring in text are characterized by being uncontrolled as to word forms, synonyms, and specificity. The only self-evident relations between words drawn from texts are those revealed by spelling. Terms assigned to text can, on the contrary, be controlled as to word forms, synonyms, and specificity. Relations between them can be displayed by coding them and in other ways.

i) Terms unrelated*

Atomizing - 10
Denitration - 2
Drying - 8
Gas - 8
Latent heat - 8
Liquid - 8
Pyrites - 4
Radiation - 8
Reaction - 10
Roasting - 4
Spraying - 8
Sulphite liquor - 6
Sulphur dioxide - 6
Suspension - 10
Uranium oxide - 2
Uranyl nitrate - 2
Wall - 8

ii) Terms interlocked

1. Atomizing - Reaction - Suspension
2. Atomizing - Drying - Gas - Latent heat - Liquid - Radiation - Spraying - Wall
3. Drying - Spraying - Sulphite liquor - Sulphur dioxide
4. Drying - Pyrites - Roasting - Spraying
5. Denitration - Uranium oxide - Uranyl nitrate

iii) Terms structurally related by roles

Sulphite liquor - A
Sulphur dioxide - B
Spraying - J
Drying - H

where

A = input
B = output
J = agency
H = operation

(iv) Terms structurally related by natural language syntax

Sulphur dioxide obtained by drying sulphite liquor by spraying

Denitration of uranyl nitrate to produce uranium oxide

* The numbers added illustrate the use of weights

Fig. 1. Organization of index record

Indexing languages - irrespective of the nature of their terms - may also differ with respect to the organization of terms in the index record. The terms may be:

- a) *unrelated* - only the joint presence of terms is recorded.
- b) *interlocked* - terms are linked into subsets, each concerned with a particular topic in or section of the item indexed.
- c) *structurally related* - the semantic relations between terms in a subset are recorded. This may be achieved by adding to each term a role or facet indicator; or by combining the terms in an ordered sequence linked by relational operators or natural language syntax, free or ruled.⁴
- d) *weighted* - according to their relative importance in the item indexed.

An example of these variations is shown in Fig. 1.

Search languages - again irrespective of the nature of their terms - can be recorded in the search record with the same variations. In addition, the form of the record can be modified by *search logic*: terms (or subsets of terms) can be related by the connectives *and*, *or*, *not*; sums of terms may also be formed by *truncation*. A search may require that terms be juxtaposed in an index record, or that they occur within a certain subset of an index record. For an example, see Fig. 2.

a) Using Boolean logic

Carbides or Carbon
and Plutonium or Uranium
or Plutonium Compounds
or Uranium Compounds
and Irradiation or Swelling
or Fission products
or Radiation effects
not Protons or Electrons
or Proton beams
or Electron beams
or Gamma radiation

d) Using index record structure

A search can require that the search terms be *interlocked* in the index record, or carry certain *roles*, etc.

b) Using weighting

Cryogenic - 5
Low temperature - 5
Metals - 1
Alloys - 1
Superalloys - 1
Superconductors - 1

c) Using truncation

* FUNG*
SPOR*
MYOE
YEAST*
MOULD*
BLIGHT*
MILDEW*
MICROORG*
MYCOS
CILLIUM
ERGILL
TORUL

Fig. 2. Organization of search statement

A *dictionary* is a list of terms in one language, giving equivalents in another. Some types of dictionary relevant to retrieval are:

- i) a list of natural language text terms with equivalents in controlled assigned terms (Fig. 3)
- ii) the reverse (Fig. 4)
- iii) a list of natural language terms (text or assigned) with code equivalents (Fig. 5)
- iv) the reverse (Fig. 6)
- v) a list of terms in one retrieval language, with equivalents in another (Fig. 7). This may be called a *concordance*.

disk hernia, cervical
cervical disk, hernia
cervical disk, hernia
intervertebral disk, hernia, cervical
intervertebral disk, cervical hernia
cervical intervertebral disk, hernia
cervical intervertebral disk, hernia
hernia, disk, cervical
hernia, cervical disk
hernia, cervical intervertebral disk
hernia, intervertebral disk, cervical
hernia, nuclei pulposi, cervical
cervical hernia nuclei pulposi
slipped disk, cervical
cervical slipped disk
disk, slipped, cervical
intervertebral disk, slipped, cervical

Fig. 3. *Excerpta Medica*: example of use of synonym thesaurus

The indexer may assign any one of the listed terms to an item, and the computer will transform it to the standard term "cervical disk hernia".

A *thesaurus* may be defined as a display of the terms in a retrieval language showing semantic relations between them. In this sense, both classified schedules and diagrammatic displays (of the Euratom type) are thesauri. The lists commonly known as thesauri combine the thesauric function (BT, NT, and RT relations) with that of a dictionary (the USE and UF links), and may be called *relational* thesauri (Fig. 8). Another type will be discussed later.

547 Indication

N. indication, pointing out, drawing attention, showing 522n. manifestation; signification, meantin....

Extract from Roget's Thesaurus....

547 Indication

N. *indication*, pointing out, drawing attention, showing 522n. *manifestation*; signification, meaning 514n. *connotation*; notification 524n. *information*; symbolization, symbolism 551n. *representation*; symbol, conventional s., x, letter; magic symbol, pentacle 983n. *talisman*; natural symbol, image, type, figure; token, emblem, figurehead (see *badge*); something to go by, symptom, sign 466n. *evidence*; tall-tale sign, bluish 526n. *disclosure*; nudge, wink 524n. *hint* (see *gesture*); straw in the wind, sign of the times 511n. *omen*; clue, scent, whiff 484n. *detector*; noise, footfall 398n. *sound*; interpretation of symptoms, symptomatology, semeiology, semeiotics 520n. *hermeneutics*; pointer, finger, forefinger, index finger (see *indicator*); indice, exponent 85n. *number*; guide, index, thumb-i, 87n. *directory*; key 520n. *interpretation*; marker, mark, guide-m.; white mark, blam; nick, scratch 260n. *notch*; line, score, stroke; note, side-n., under-runner, catchword (see *punctuation*); stamp, print, impression; stigma, stigmata;

prick, tattooing, tattoo-mark 263n. *perforation*; mole, scar, birth mark, strawberry mark 845n. *blemish*; legend, caption 590n. *description*; inscription, epitaph; motto, cipher; love-token, favour (see *badge*).
identification, naming 561n. *nomenclature*; 77n. *classification*; means of identification, brand, trade-mark, imprint (see *label*); name and address; autograph, signature, hand 586n. *script*; finger-print, footprint, spoor, track, trail 548n. *trace*; secret sign, password, open sesame, watchword, countersign; diagnostic, markings, stripes, spots, colour, colouring, cloven hoof; characteristic, trait, lineament, outline, form, shape 445n. *feature*; personal characteristic, trick, trick of speech, shibboleth; mole, scar, birth-mark, strawberry mark 845n. *blemish*; divining rod 484n. *detector*; criterion 461n. *testing agent*; mark, note.
symbolology, symbolization, dactylogy; cipher, code 525n. *secrecy*; symbolography, picture-writing, hieroglyphics 586n. *script*; gypsy signs, hobo s., scout s.
gesture, gesticulation, sign-language, semeiology, dactylogy; deaf-and-dumb language; sign 524n. *hint*; pantomime, by-play, dumb-show, charade, dumb c.; demeanour, look in one's eyes, tone of one's voice 445n. *mien*; motion, move; tick, twitch 318n. *spasm*; shrug, shrug of the shoulders; wag of the head, nod, beck, wink, flicker of the eyelash, twinkle, glance, ogle, leer, grimace 438n. *look*; smile, laugh 835n. *laughter*; touch, kick, nudge, jog, dig in the ribs 279n. *knock*; hug, clap on the shoulders; hand-pressure, squeeze of the hand, handshake, grip 778n. *retention*; push, shove 279n. *impulse*; pointing, signal, waving, wave, hand-signal, wave of the hand; raising one's hand, wagging one's forefinger; drumming one's fingers, tapping one's foot, stamp of the foot 822n. *excitable state*; clenching one's teeth, gritting one's s. 599n. *resolution*; gnashing one's teeth, snap, snapping one's jaws 892n. *irascibility*; wringing one's hands, tearing one's hair 836n. *lamentation*; clenched fist 711n. *defiance*; flag-waving, umbrella w., hat-w. 876n. *celebration*; clap, clapping, hand-c., cheer 923n. *applause*; hiss, hissing, hooting, boo, booing, Bronx

FIG. 4. An extract from Roget's Thesaurus, showing assigned concepts (such as *Indication*) and natural language words

Peruvian woolly mouse 636.932.44

Pests

farm animals

prevention and control 636.093

fish 639.2.09

honeybee 638.157

silkworm 638.257

Pets 636.045

Pheasants 636.594

Physiological changes due to domestication 636.068

Pigeons 636.596

Pigs 636.4

Polecat 636.934.56

Pollen 638.138

Pollination by honeybees 638.19

Ponies 636.16

Porpoise hunting 639.254.3

Poultry 636.5 and 637.54

Predators of farm animals 636.092.2/3

Preparation

feeding-stuffs 636.085.6

Fig. 5. Index to UDC schedule

CODE	CAPTION	ULT
581	Respiratory System	
05	Function, Biochemistry	
200	Clearance, Ciliary Action	L-Resp-Clear
300	Control and Regulation	L-Resp-Control
400	Diffusion of Gases	L-Resp-Diffus
600	Mechanics, Compliance	L-Resp-Mech
800	Surface Tension	L-Resp-Surf
995	Other	L-Resp-Funct-O
997	CMC	L-Resp-Funct-CMC
30	Diseases and Conditions	
050	Abscess	
075	Atelectasis	L-Resp-Abscess
100	Asthma	L-Atelect
50	Intractable	L-Resp-Asthma-Intract
95	Other	L-Resp-Asthma-O
200	Bronchitis, Acute	L-Bronchit
300	Chronic Respiratory	
20	Bronchiectasis	L-Bronchiect
40	Emphysema	L-Emphysema
97	CMC	L-Chronic-Resp-CMC
400	Hyaline Membrane	L-Hyaline
500	Irritants, Inhaled	
50	Gases	L-Inhal-Gas
60	Particulates	L-Inhal-Part
95	Other	L-Inhal-O
700	Pneumonia, Pneumonitis	L-Pneumonia
725	Pulmonary Alveolar	L-Proteinosis
750	Proteinosis	L-Pul-Edema
775	Pulmonary Edema	L-Pul-Fib
995	Pulmonary Fibrosis	L-Resp-Dis-O
997	Other	L-Resp-Dis-CMC
	CMC	

Fig. 6. Science Information Exchange: extract from classified thesaurus

The indexer assigns one or more specific term (from the column labelled ULT) to each item, e.g. Lresp-Asthma-Intract. The computer scans the index entries, and translates each into its code, in this case 581 30 100 50. The item will be selected by a search at any level of the hierarchy (such as 581 30 100 Asthma, or 581 Respiratory system).

621.791	H	H	621.791
621.791(204)	IN	HAK	621.791.05
621.791.33	WAF	HAD	621.791.55
621.791.37	FK	HA	621.791.55
621.791.38	WA	HAd	621.791.555-52
621.791.542.1	FN	HB	621.791.556
621.791.620.1	FJ	HC	621.791.559
621.791.657.478.8	WD	HEJ	621.791.75
621.791.669.1	K	HEF	621.791.75.042
621.791.002	WE	HE	621.791.75.042-51
621.791.01	RLT	HEd	621.791.756.042-51
621.791.01.669	HCH	HEf	621.791.85.042-51

Fig. 7. Concordance between the IIS and UDC systems of classification

CONVERGENCE		COOLANTS	
RT	DIVERGENCE	RT	AIR CONDITIONING
	FOURIER ANALYSIS		AIR COOLING
	FUNCTIONAL ANALYSIS		BRINES
	INTEGRALS		COOLERS
	SEQUENCES		COOLING
	SERIES		COOLING SYSTEMS
	VARIABILITY		CUTTING FLUIDS
			DRY ICE
			ICE
			LIQUID COOLING
			REFRIGERANTS
			WATER COOLING
CONVERSION		COOLERS	
RT	CONVERTERS	UF	AFTERCOOLERS
	CRACKING PROCESS		INTERCOOLERS
	DIGESTION	NT	BEVERAGE COOLERS
	EXCHANGING		MILK COOLERS
	LIQUEFACTION		OIL COOLERS
	REACTIONS (CHEMICAL)		UNIT COOLERS
	REFINING		WATER COOLERS
	REVISIONS	RT	AIR CONDITIONERS &
			AIR CONDITIONING EQUIPMENT
CONVERSION COATING (PROCESS)			COOLANTS
BT	COATING		COOLING
RT	CONVERSION COATINGS		COOLING SYSTEMS
	DIFFUSION COATING (PROCESS)		DEHUMIDIFIERS
	HARD SURFACING		FREEZERS #
			ICE REFRIGERATION
			REFRIGERATION MACHINERY &
			REFRIGERATION
			REFRIGERATORS &
CONVERSION COATINGS			
BT	COATINGS		
RT	CONVERSION COATING (PROCESS)		
	FINISHES		
	PROTECTIVE COATINGS		
CONVERSION TABLES			
USE	TABLES (DATA)		
CONVERTERS		COOLING	
	(USE MORE SPECIFIC TERM IF POSSIBLE)	UF	CHILLING
RT	AIR-BLOWN CONVERTERS		HEAT DISSIPATION
	BOTTOM-BLOWN CONVERTERS	NT	AIR COOLING
	BREEDER REACTORS		CONDUCTION COOLING
	CATALYTIC CONVERTERS		CONVECTION COOLING
	CONVERSION		EVAPORATIVE COOLING
	COPPER CONVERTERS		EXPANSION COOLING
	DATA CONVERTERS		FILM COOLING
	ELECTRIC CONVERTERS &		HOT QUENCHING &
	ELECTRIC GENERATORS &		LIQUID COOLING
	IMAGE CONVERTERS		MAGNETIC COOLING
	NUCLEAR CONVERTERS		QUENCH AGING #
	NUCLEAR POWER REACTORS		QUENCHING (COOLING) &
	NUCLEAR PRODUCTION REACTORS		RADIANT COOLING

Fig. 8. Extract from Engineering Thesaurus

BEST COPY AVAILABLE

FUNCTIONAL PURPOSES

When the functions of specific reference and generic survey were first stressed, advocates of each tended to denigrate the other. Many were the battles fought between abecedarians and classifiers. Each approach was acclaimed as the best, suitable in all circumstances. Even the symbiotic approach of Ranganathan suggested that a certain combination of alphabetic index and classification was the best. This search for a do-all language, to serve all purposes, is by no means over, though nowadays the battle is often between the advocates of natural and of controlled languages. It is as though carpenters were arguing as to which tool could be used for all jobs - the saw, the chisel, the hammer, the plane, or whatever.

What we need is a more detailed spelling out of the varied functions of retrieval language, and studies relating structure to function. Some aspects of this problem will now be considered. All the following functions can be found in one retrieval system or another, and there may well be others. A retrieval language may:

- 1) provide for the selection of bibliographic records with unique characteristics - for example, any record containing a specified string of juxtaposed words, such as "enzyme activity in mammalian cells".
- 2) provide for the selection of a set of records likely to be relevant to a particular topic - the function of specific reference.
- 3) provide for the selection of a group of records lying within a certain subject field - for example, in an agricultural and biological index, all items concerned with Parasitism. This is the function of generic survey.
- 4) provide for the sequencing of a set of selected records according to probable relevance to a particular topic - by means of some ranking device.
- 5) provide for the ordering of a group of records in a field into a meaningful sequence for example, to arrange by subject the selected items concerned with Parasitism.
- 6) provide for the automatic conversion of index entries from one form to another - an example is the truly syntax developed by Lynch, whereby rotated articulated index entries can be derived; an earlier example is the chain procedure of Ranganathan, which imposes restrictions on both the class numbers and the verbal terms used.
- 7) provide for conversion from one language to another - as far as terms are concerned, dictionaries do this; the interconvertibility of structural features in general is discussed in ref. 5.
- 8) Give aid to the searcher in his choice of search terms - this is the function of thesauri.

No retrieval system provides all the facilities, because none attempts to serve all these purposes. There is no single type of retrieval language that can "do all", that is "the best". User needs determine what functions should be provided, and different functions require different structures.

STRUCTURE AND FUNCTION

Function 1 makes the least demands on structure. The searcher is simply asking 'is there a record containing the following term or string of terms?'

The usual example is in whole-text searching, when the exact wording of the original text is the target of interest. Any control of the text words will then worsen performance. Natural language words and syntax is the required structure, for both indexing and search.

Function 2 is today a battleground for the claims of rival retrieval languages. Both text terms and assigned terms can be and are used in specific reference retrieval systems. Even if we leave to one side evidence presented by the Cranfield project,⁶ the modes of use encountered in operational systems make one thing clear: if word forms and synonyms are not controlled in the indexing language, then for adequate performance, they must be summed in the search language. The examples of Fig. 2 illustrate this. Since summation implies multiple search, it is inconvenient for visually scanned systems. Uncontrolled text words and summation are suitable only for machine search.

The devices just discussed serve to increase recall. There is less evidence of the value of devices to increase precision. Interlocking of various kinds (splitting an item into sections or topics; the use of

links; search "within barriers", e.g. within the limits of a sentence in a natural language abstract) is widely practised in both visual and machine systems, and must reduce irrelevant search output. Weighting index terms (which divides them among major and minor topics in the item) can have the same effect.

To structure entries in a visually searched, pre-coordinated index (whether alphabetical or classified) produces a more easily used tool, in which the human searcher is less likely to lose his way. But, given exhaustive search (such as a machine performs), there is little evidence that structural relations affect performance favourably: improvement in precision is countered by loss in recall.

Function 3 is the provision of generic survey facilities. By definition, this requires the selection of records of various levels of specificity, concerning many varied topics that lie within the field surveyed. These can be selected by two techniques: either by the *ad hoc* summation of terms or topics that collectively cover the field; or by the use of an indexing language that relates specific to generic terms.

The first technique is simply an extension of the summation already illustrated (Fig. 2). *Ad hoc* summation may be aided by any thesaurus of the kinds already discussed. The second technique is that for which hierarchical classification is especially devised. Fig. 9 illustrates some of the many specific topics that will be brought together in a search for class 8, Biophysics. We are left with the conclusion: for generic survey, if generic relations have not been built into the indexing language, they must be provided by summation in the search language. It is particularly necessary to build in generic relations for visually searched systems.

June 1971		RETRIEVAL LANGUAGES
.8	<u>BIOPHYSICS</u>	
.81	PHYSICAL PROPERTIES OF BIOLOGICAL ENTITIES (for electrical and magnetic properties of biological entities combine with .23 section) (for optical properties combine with .24 section) (for chemical reactions in biological entities combine with .49 section) (for thermal and thermodynamic properties of biological entities combine with .51 section) (for mechanical and rheological properties of biological entities, including acoustical properties, combine with .52 section)	
.811	Mass, density	
.812	Size, shape, structure	
.813	Chemical composition	
.82	SENSORY PHENOMENA	
.821	Vision, visual perception	
.8211	Visual threshold	
.8212	Color perception	
.8213	Depth perception	
.8214	Defective vision	
.822	Hearing, auditory perception	
.8221	Auditory threshold	
.8222	Pitch perception	
.8223	Binaural hearing	
.8224	Defective hearing	

FIG. 9. American Institute of Physics: extract from classified thesaurus

The indexer assigns one or more specific codes to each item, e.g. 8214 Defective vision. The item will be selected by a search at any level of the hierarchy (such as 821 Vision, or 8 Biophysics)

Fig. 9. American Institute of Physics: extract from classified thesaurus

There is a weakness in hierarchical classification. Often each specific term occurs in only one hierarchy. If the classification is of the enumerative pre-coordinated type, each specific topic may occur in only one hierarchy. This is overcome by faceted classification, where a compound class number is hierarchically subordinate to each facet or subfacet in it, but there may still be a restriction on the number of generic terms linked to each specific term. One can readily visualize a retrieval language of the form shown in Fig. 10, in which each listed term is tagged by every (coded) generic term that is related to it. These can be not only facet and subfacet codes, but also codes for fields. Such a list may be called a *generic* thesaurus.

BEST COPY AVAILABLE

Function 4 requires a ranking of records. Ranking methods in use depend either on the *number* of terms in index and search records, or on *weights* assigned to them. Any retrieval language serving this function must therefore provide a record that can be separated into individual units (terms or other units) that can be counted and/or weighted. With this proviso search output using any retrieval language can be ranked. If the output is consistently large, there is obvious value in the technique.

Function 5 is concerned with meaningfully ordering a group selection. If the main file, from which selection is made, is already in orderly sequence (as in alphabetic or classified pre-coordinate index) then a selected group can be similarly ordered. But if the main file is in a sequence not determined by subject, then the search must specify relevant codes, terms, or topics on which a meaningful arrangement can be made. To take a simple example from notched cards: select the group notched in hole A; from it, select those notched in hole B, and list these; then from the group select those notched in hole C, and list these; and so on. This procedure is independent of the retrieval language used it provides an output that is more adapted to human scanning.

camel	TYL	ART	MAM	DOM	FUR	
cat	FEL	CAR	MAM	DOM		
cattle	BOV	ART	MAM	DOM	LEA	MEA
chicken	GAL	RAS	AVE	DOM	MEA	
chinchilla	HIS	ROD	MAM	FUR		
deer	CER	ART	MAM	MEA	LEA	GAM
dog	CYN	CAR	MAM	DOM		
duck	ANS	PAL	AVE	DOM	MEA	
fox	CYN	CAR	MAM	FUR		
goat	BOV	ART	MAM	DOM	LEA	
goose	ANS	PAL	AVE	DOM	MEA	
grouse	GAL	RAS	AVE	MEA	GAM	
horse	EQU	PER	MAM	DOM		
mink	ARC	CAR	MAM	FUR		
pheasant	GAL	RAS	AVE	MEA	GAM	
pig	SUO	ART	MAM	DOM	MEA	LEA
rabbit	LAG	ROD	MAM	MEA	FUR	GAM
seal	PIN	CAR	MAM	FUR		
sheep	BOV	ART	MAM	DOM	MEA	LEA
turkey	GAL	RAS	AVE	MEA		

ANS	Anseriformes	GAM	Game animal
ARC	Arctoidea	HIS	Histricomorpha
ART	Artiodactyla	LAG	Lagomorpha
AVE	Aves	LEA	Leather-providing
BOV	Bovoidea	MAM	Mammalia
CAR	Carnivora	MEA	Meat-providing
CER	Cervoidea	PAL	Palmipedes
CYN	Cynoidea	PER	Perissodactyla
DOM	Domesticated	PIN	Pinnipedia
EQU	Equines	RAS	Rasores
FEL	Feloidea	ROD	Rodentia
FUR	Fur-providing	SUO	Suoidea
GAL	Galliformes	TYL	Tylopoda

FIG. 10. The form of a generic thesaurus. The codes represent terms generic to the listed terms

Fig. 10. The form of a generic thesaurus. The codes represent terms generic to the listed terms

Function 6 can appear only if the terms in index records are structurally related - essentially it is concerned with various methods of rotating an ordered sequence of terms. One of the examples already quoted (Lynch) is wholly a matter of ruly rotation, as are comparable techniques developed by Coates, Austin, and indeed KWIC indexing. Chain procedure can involve a dictionary conversion, as well as rotation. The purpose of these rotations is to provide multiple access for visual scanning, in order to improve the chance of high recall.

Function 7, conversion of terms from one retrieval language to another, appears in several contexts. It is essentially a device to reduce the intellectual labour of using two index languages, by providing a clerical conversion between them. As already illustrated, a dictionary can be used to confound word forms and synonyms (Fig. 3), to assign code numbers (Fig. 5) or the codes of generic terms (Fig. 10), and to switch between languages (Fig. 7).

Finally, we come to function 8. Relational thesauri continue to proliferate. Undoubtedly they provide first aid in searching a field not intimately familiar to the searcher. But that the benefit received is always commensurate with the construction effort involved has yet to be demonstrated. Thesauri are used (i) to provide alternative search paths, and (ii) to choose terms for summation. There is evidence (a) that a previously constructed thesaurus rarely sets out all the alternatives that might helpfully be chosen, so it must be supplemented by other aide; and (b) that purely mechanical summation of the related terms displayed is likely to produce irrelevant output.

This discussion of structure and function in retrieval languages is tentatively summed up in the following tables.

<i>Function</i>	<i>Needed elements of retrieval language</i>
1 String selection	Natural language words and syntax.
2 Specific reference	(a) Either natural language terms and summation or word form and synonym control (b) interlocking. (c) Structural relations (for visual search).
3 Generic survey	(a) Either specific terms and summation or generic relations (hierarchy or thesaurus).
4 Relevance ranking	Separately countable and/or weightable terms.
5 Subject ordering	Either an ordered main file or specified ordering at output
6 Entry conversion	(a) Structural relations (b) Rotation rules
7 Language conversion	(a) Dictionary. (b) See ref. 5.
8 Search aid	Thesaurus
<i>Language elements</i>	<i>Comments</i>
Natural language	Only suitable for machine retrieval, because of summation difficulties
Control of word forms and synonyms	Essential for effective visual retrieval, can be helpful for machine retrieval
Control of specificity	Dependent on user approach
Interlocking	Useful in all forms of retrieval

Structural relations in index record	Helpful in visual search, inadequate evidence of value in machine retrieval
Generic relations between index terms	Necessary for generic visual search, can be helpful for machine retrieval
Weighting	Valuable mainly for ranking search output, useful only in machine retrieval
Dictionaries	Can be used either in visual or machine systems.
Thesauri	Of some value in formulating search programmes.

REFERENCES

1. Vickery, B.C. *On retrieval system theory*. 2nd ed. London, Butterworths, 1965
2. Coyaud, M. *Introduction à l'étude des langages documentaires*. Paris, Klincksieck, 1996.
3. Meadow, C.T. *The analysis of information systems*. New York, Wiley, 1967.
4. Systems using an ordered sequence of terms in an index record include the analets of Farradane, the SYNTOL of Gardin, the ruly combinations of Coates, and the articulated entries of Lynch.
5. Vickery, B.C. Coding for interconvertibility. *Advances in documentation and library science*, vol. 3, part 2, pp. 1159-74. New York, Interscience, 1960.
6. Cleverdon, C.W. The Cranfield tests on index language devices. *Aslib Proceedings* 1967, vol. 19, pp. 173-94.

VICKERY: STRUCTURE... This paper first appeared in *Journal of Documentation* Vol. 27, No. 2, 1971 p. 69-82 (ISSN 0022 0418) and has been reproduced here with the kind permission of the original publisher, ASLIB, the Association for Information Management, Information House, 20-24 Old Street, London EC1V 9AP, United Kingdom. **VICKERY:**

KNOWLEDGE REPRESENTATION: A BRIEF REVIEW

B.C. Vickery

University College, London

The problems of representing knowledge in computer systems are common to information science, artificial intelligence, psychology and linguistics. The paper offers a brief review of the structures and techniques that have been developed in these different disciplines. It looks at the semantic structure of sentences, at roles, categories and relations in subject analysis, at semantic primitives, and at knowledge representation for reasoning. The paper concludes with a note on a prototype expert system that makes use of some of these techniques.

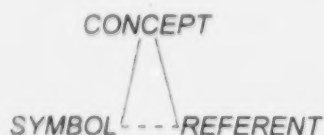
'KNOWLEDGE', SAID SAMUEL JOHNSON, 'is of two kinds. We know a subject ourselves or we know where we can find information upon it.' In this way we may distinguish between personal and public knowledge. Documentation is particularly concerned with the second kind of knowledge, and agrees with Goethe: 'all our knowledge is symbolic'.

The representation of knowledge in symbolic form is a matter that has preoccupied the world of documentation since its origin. The problem is now relevant in many situations other than documents and indexes. The structure of records and files in databases; data structures in computer programming; the syntactic and semantic structure of natural language; knowledge representation in artificial intelligence; models of human memory: in all these fields it is necessary to decide how knowledge may be represented so that the representations may be manipulated.

Many different techniques of representing knowledge have been developed in each field, the variations often being due to the different kinds of manipulation that are to be performed. But there are some interesting convergences, and it seems likely that each field may have something to learn from the others. This paper aims to review different forms of knowledge representation, seeking insights that may be applicable to our own field of documentation (information science).

SYMBOLS AND SYMBOL STRUCTURES

The use made here of the expression 'symbol' may be illustrated by the familiar diagram:



Corresponding to a referent in the outside world (a real entity) or in an imagined world within public knowledge (e. g. a unicorn), a concept exists within personal knowledge structures, in one or more minds. The concept can be represented by a symbol, such as a word or string of words, that may be

uttered or recorded (e.g. written). The symbol thus indirectly represents the referent. The same referent may give rise to varied concepts in different minds. The same concept may be represented by different symbols.

Knowledge can be represented by symbols combined in various ways, the most familiar of which is natural language: 'cats eat fish' is a sentence that represents a meaningful combination of symbols for three concepts. The expression 'symbol structures' relates to the: Patterns and rules used to control both the structure of individual symbols (e.g. the rules for plurals) and the structure of symbol combinations (e.g. natural language syntax).

Natural language is a symbolic medium that is used to express any and every kind of knowledge, for any and every purpose. The unitary forms of symbol in natural language are the word (the 'lexeme') and the 'morpheme' - the smallest component of a word that has meaning, such as the hyphenated parts of the word 'un-expect-ed-ly'. The symbol structure of natural language is concerned with rules relating to word form and word order. Traditional grammar categorises words as nouns, verbs, adjectives, prepositions, adverbs, conjunctions, articles, and so on, and provides rules about category sequence. Such rules are used in the symbol manipulation process known as parsing, which takes a given utterance or text and fits it to one of the allowable category sequences. This is a necessary step in either human or computer understanding of natural language.

Grammar, natural language syntax, is mainly concerned with the *usage* of language, with the ways in which a given combination of concepts may conventionally be expressed. As linguistic behaviour changes, so do the allowable category sequences. The same logical statement can be expressed in different ways. In this paper the main focus of interest will not be on this, but on the different *logical* patterns in which knowledge may be represented.

Let us begin by considering the kinds of formalism that have been used to structure strings of words such as sentences and subject statements.

SEMANTIC STRUCTURE OF SIMPLE SENTENCES

A simple declarative sentence is a grammatically complete string of words expressing a single thought, for example:

1. London is in England
2. Edinburgh is between Aberdeen and York
3. Mickey Mouse squeaks
4. The moon is silver
5. My brother is swimming the Channel
6. I am hungry
7. Tarzan loves Jane
8. Bill gave Mary the book
9. This picture is similar to that picture
10. John is a teacher
11. Harry is the brother of Alice
12. Robert is my father

The simplest semantic analysis of such sentences uses the word categories of predicate and argument (see, for example, Hurford and Heasley¹). In each of the above examples, there are words that have obvious referents in the real world or public knowledge (London, England, Mickey Mouse, the moon, my brother, Tarzan, Bill, Mary, this picture, etc.). Each of these is thus a referring expression. Within each sentence there is a further word that is not necessarily a referring expression: it is called a 'predicate'. The predicates in the examples include: in, between, squeaks, silver, swimming, hungry, loves, gave, similar, teacher, brother. The referring expressions are known as 'arguments'. So, for some of the examples, we have:

1. Predicate = in, arguments = London, England
4. Predicate = silver, argument = the moon
7. Predicate = loves, arguments = Tarzan, Jane

The predicates can be verbs (squeaks, loves), adjectives (silver, hungry), nouns (teacher, brother), prepositions (in, between). A predicate can have a single argument (John is a teacher, the moon is silver), two arguments (London is in England, Harry is the brother of Alice), or sometimes three (Bill gave Mary the book). In example 12, both 'Robert' and 'my father' are referring expressions. The remaining word, 'is', is the predicate, and is said to express the identity relation.

A simple sentence can thus be represented by two basic semantic components - a predicate and one or more arguments (referring expressions). The function of the predicate is to specify the logical relationship(s) of the argument(s), i.e. how they participate in the situation represented by the sentence. More precise specification of relationships is used in case grammar (e.g. Fillmore²). The 'case' or role of each argument is specified. The approach has been particularly useful where the predicate is a verb, using roles such as Agent, Patient, Instrument, Beneficiary, Location. Examples are:

1. John opened the door with the key
AGENT (PREDICATE) PATIENT INSTRUMENT
2. Keith gave Joan the chocolates
AGENT (PREDICATE) BENEFICIARY PATIENT
3. Bill travelled from London to Luton
AGENT (PREDICATE) LOCATION LOCATION

Each verb may be annotated to show which role(s) must obligatorily be associated with it, and which can be optionally associated. Thus (putting optional roles in brackets), we have:

open: (AGENT) PATIENT(INSTRUMENT)
put: (AGENT) PATIENT (INSTRUMENT) LOCATION
give: (AGENT) PATIENT BENEFICIARY

Other roles that have been used are Object (usually the entity that is the central focus of the sentence), Recipient (often the same as Patient), Coagent, and Time. The use of roles has not been clarified for sentences in which the predicate is a noun, adjective, preposition or the identity relation.

The same idea of roles has been introduced by Lindsay and Norman³ in their modelling of human memory structures. Their model includes the use of roles that characterise the 'parts of an event', for example:

Action	The event itself. In a sentence, the action is usually described by a verb: The man was <i>bitten</i> by the dog.
Agent	The actor who has caused the action to take place: The man was bitten by the <i>dog</i> .
Conditional	A logical condition that exists between two events: A shark is dangerous <i>only if</i> it is hungry. John flunked the test <i>because</i> he always sleeps in lecturer.
Instrument	The thing or device that caused or implemented the event: The <i>wind</i> demolished the house.
Location	The place where the event takes place. Often two different locations are involved, one at the start of the event and one at the conclusion. These are identified as from and to locations: They hitchhiked <i>from</i> La Jolla <i>to</i> Del Mar.
Object	The thing that is affected by the action: The wind demolished the <i>house</i> .
Purpose	Identifies the purpose of the event: Jack took Henry to the bar <i>to get him drunk</i> .

BEST COPY AVAILABLE

Quality	A descriptor, one that modifies a concept: The surf was <i>heavy</i> . There were 93 people in class.
Recipient	The person who is the receiver of the effect of the action: The crazy professor threw the book at <i>Peter</i> .
Time	When an event takes place: The surf was up <i>yesterday</i> .

ROLES, CATEGORIES AND RELATIONS IN SUBJECT STATEMENTS

Roles have been used in essentially the same fashion in information retrieval, one being assigned to each individual component of a subject statement to indicate how it participates in the topic represented by the statement. For example, Costello and Wall⁴ used the following role indicators in chemical technology:

- Uses of (for), applications of (for), used to (for)
- Causes, influences, independent (controllable) variables
- Reactant, input, raw material
- Special agent
- Medium, vehicle, solvent
- Byproduct, waste, scrap, contaminant
- Product output; manufacture, production, fabrication or synthesis of
- Research on, development of, investigation of
- Dependent (affected) variable
- Design of
- Physically processed, treated, changed, handled; passive

There have been two other approaches in information retrieval to the structuring of subject statements. The first has been, within a given subject field, to assign each term to a semantic category or 'facet'. Within science and technology, the following facets have been found useful.³

Categories	Example
Things, substances, entities	
Naturally occurring	Minerals, Plants, Soils
Products	Bridges, Engines, Fibres
Mental constructs	Equations, Rectangles
Their parts	Beam, Wheel, Wing
Constituents, materials	Metal, Glass, Nitrogen
Organs	Heart, Seed
System of things	
Attribute of things	
Qualities, properties, including	Cohesion, Colour, Solubility
Structure	Layers, Profile
Measures	pH, Ampere
Processes, behaviour	Vibration, Inflammation
Object of action (patient)	Any thing can be a patient
Relations between things, interactions	
Effects	Inhibition, Stimulation
Reactions	Nitration, Symbiosis
Operations on things	
Experimental	Cutting, Breeding
Mental	Calculation, Reasoning
Operating agents	Any thing can act as an agent
Properties of attributes,	
relations and operations	Rate, Variation
Operations on attributes,	
relations and operations	Measurement, Control
Place, condition	
Time	

The other approach to subject structure has been to interpose, between each pair of successive components, an operator signifying the relation subsisting between them. For example, Farradane⁶ used the operators Concurrence, Comparison, Association, Equivalence, State, Appurtenance, Distinctness, Reaction and Causation. Other examples may be found in books by Vickery⁷ and de Grolier.⁸

REPRESENTATION OF KNOWLEDGE IN CLASSIFICATION

The ideas of facets and roles have been brought together in a wider context in subject classifications. The structure of a fully developed classification can be technically described in the following way:

1. A whole collection (universe) of entities (in this case subject terms) may be divided into a series of *fields* (for example, Physics into Mechanics, Optics, Heat, Electricity, etc). Each field is therefore a broad grouping. The fields may be arranged in some meaningful sequence.
2. Each field may be divided into a series of *facets* (for example, Chemistry into Substances, Properties, Reactions, etc). Each facet is therefore a grouping. The facets within each field may be arranged in some meaningful sequence.
3. Each facet may be structured into a *hierarchy*, subdivided stage by stage by the application of a series of *characteristics*. These are applied in an ordered sequence. Any sequence of terms generated by successive subdivisions is a *chain* (for example, the chain 'Machines - Mechanical and thermal machines - Nuclear reactors - Heterogeneous reactors - Thermal heterogeneous reactors - Graphite moderated - Gas cooled'). Any one level of subdivision gives rise to a group of terms that constitute an *array* (for example, within the class Metals the array 'Bismuth, Lithium, Mercury, Potassium, Sodium'). The terms in each array may be arranged in some meaningful sequence.

4. Rules may be provided for combining (co-ordinating) terms from the same array, from different arrays in the same facet, from different facets in the same field, and from different fields. These rules may involve the use of special relational operators or role indicators.
5. Each field, facet and term may be coded to fix its position in the whole system, and to facilitate unambiguous combination with other codes.
6. An alphabetical index to the terms, facets and fields may be provided, showing the code of each.

RELATIONS IN THESAURI

Standard lists of terms used for subject indexing and reprieve have long been in existence. Relations between words were at first confined to 'see' references from term A to term B, implying that B should be used instead of A, being a preferred synonym or near-synonym; and 'see also' references, implying that the terms were semantically related in some unspecified way. The modern thesaurus separates out from the 'see also' references the relationship of class membership, and links a class to its members as a 'broader term' to a set of 'narrower terms'. All other 'see also' references are treated as 'related terms'.

Willels,⁹ in a study of term relations in thesauri, identified a number of different kinds of 'related term', for example:

- An entity and its parts
- An entity and its properties
- An entity and a process
- The use of an entity
- A property and its measure
- A process and its property
- An operation and a device that carries it out
- A product and the device that produces it
- A cause and its effect

SEMANTIC PRIMITIVES

In all the above examples, the words in sentences, statements or indexes are used more or less as given. Another approach to knowledge representation is to reduce the given words to semantically more primitive units. An example of this in information science is the 'semantic factoring' system developed by Perry and Kent.¹⁰

They established a list of over 200 primitive semantic factors, which were used to encode about 8,000 English terms, mostly within science and technology. Each factor was represented by a four letter code, e.g. M-TN for the factor 'motion'. An individual factor could be modified in three ways:

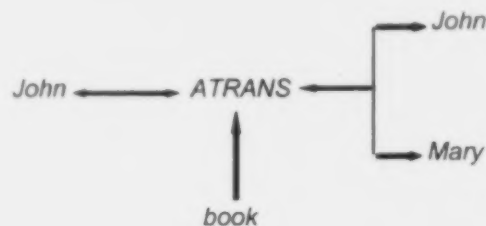
- (a) By including an 'analytic relation' code as the second letter. The analytic relations used were:
 - A Categorical: the coded term is a species of the genus denoted by the factor.
 - E Intrinsic: the coded term is something of which the factor is a component.
 - I Inclusive: the coded term is a component of the factor.
 - O Comprehensive: the coded term is made up of several members of the class represented by the factor.
 - U Productive: the coded term helps to produce the factor.
 - Q Affective: the coded term makes use of, is determined by, or is influenced by the factor.
 - W Instrumental: the coded term is produced by, acts on or is acted on by the factor.
 - X Negative: the coded term has a characteristic whose absence is denoted by the factor.
 - Y Attributive: the coded term is characterised by the factor.
 - Z Simulative: the coded term resembles, but is not, the factor.

- (b) By adding a numerical suffix to the code, thus Propulsion was coded MUTN.015.
- (c) By combining codes, thus Swimming was coded MUTN. HWDR.001, combining Motion with H-DR, Water.

Working in artificial intelligence, Schank¹¹ has developed the use of a set of four categories of concepts in a sentence, one of which is a set of primitive actions. The categories are Actions, Objects, Modifiers of actions, Modifiers of objects. The primitive actions are:

ATRANS	Transfer of an abstract relationship (e.g. give)
PTRANS	Transfer of the physical location of an object (e.g. go)
PROPEL	Application of physical force to an object (e.g. push)
MOVE	Movement of a body part by its owner (e.g. kick)
GRASP	Grasping of an object by an actor (e.g. throw)
INGEST	Ingesting of an object by an animal (e.g. eat)
EXPEL	Expulsion of something from the body of an animal (e.g. cry)
MTRANS	Transfer of mental information (e.g. tell)
MBUILD	Building new information out of old (e.g. decide)
SPEAK	Producing of sounds (e.g. say)
ATTEND	Focusing of a sense organ toward a stimulus (e.g. listen)

Around these primitives are built 'conceptual dependency' structures, e.g. 'John takes the book from Mary' is expressed as:



For the purposes of linguistic analysis, Wilks¹² has developed a set of up to 100 primitive semantic elements to express the entities, states, qualities and actions about which we speak and write. He groups them into five classes, shown below (the element is in upper case; an explanation of it in brackets).

- (a) Entities. MAN (human being), STUFF (substances), SIGN (verbal and written symbols), THING (physical Object), PART (parts of things), FOLK (human groups), ACT (acts), STATE (states of existence), BEAST (animals) etc.
- (b) Actions. FORCE (compels), CAUSE (causes to happen), FLOW (moving as liquids do), PICK(choosing), BE (exists) etc.
- (c) Type indicators. KIND (being a quality), HOW (being a type of action)
- (d) Sorts. CONT (being a container), GOOD (being morally acceptable), THRU (being an aperture) etc.
- (e) Cases. TO (direction), SOUR (source), GOAL (goal or end), LOCA (location), SUBJ (actor or agent), OBJE (patient of action), IN (containment), POSS (possessed by) etc.

Wilks uses these primitives as the element of a micro-language in which more complex concepts may be expressed.

CONNECTIONS AND RELATIONS BETWEEN SIMPLE STATEMENTS

Both logic and language provide ways of joining simple statements to form compound or complex

statements. The operators AND, OR, Nor provide the first formalism of this kind. We can state that 'Tarzan loves Jane AND Jane loves Tarzan', or 'I am hungry AND I am thirsty' (which can be elided to 'I am hungry AND thirsty' or 'Mary offered John tea or coffee', or Claire is married to Bill AND NOT to Dave'. The use of these operators in information retrieval is too well known to need an example.

The next logical operator is that of implication, expressed in English by the use of the words 'if, then' to form a conditional statement. Thus, 'IF the knight takes the bishop THEN Joe will lose the game of chess'. More complex conditionals are created by using implication operators to link compound statements that use AND/OR, for example:

1. IF the cricket ball hits the pad
AND the ball would otherwise have hit the wicket
THEN the batsman is declared 'out'
2. IF the passenger is of pensionable age
OR the passenger is a babe in arms
THEN the passenger travels free on the bus

Implicitly, computer information retrieval uses such conditionals, because the retrieval process is of the form 'IF the document subject statement includes term A AND term B OR term C, THEN select the document'.

Explicit use is made of conditional statements in computer-based expert systems that give consultancy advice derived from the specialised knowledge embodied in them. This knowledge is often represented as a set of 'production rules'.

An example from a medical consultation system, MYCIN,¹³ is as follows:

IF the gram stain of the organism is gramneg
AND the morphology of the organism is rod
AND the aerobicity of the organism is anaerobic
THEN there is suggestive evidence (0.6) that the organism is Bacteroides

Such a rule represents the medical knowledge that certain observed characteristics of a microorganism are evidence (though not conclusive proof) of its identity.

KNOWLEDGE REPRESENTATION FOR REASONING

The conditional statements in MYCIN are used to make an inference, to reason from the IF conditions to the THEN consequence. A form of knowledge representation explicitly designed for reasoning is predicate logic (see, for example, texts by Winston¹⁴ or Rich⁵). This starts with the categories noted earlier, predicate and argument, and uses the formalism 'Predicate (Argument(s))'. Thus, 'The moon is silver' is represented as 'Silver (Moon)'. Two or more such 'well-formed formulae' can be combined, thus:

Flies (Albatross) & Lays eggs (Albatross)

The ampersand is used to represent AND, the V sign to represent OR, and the - to represent NOT. The conditional 'IF an albatross flies AND an albatross lays eggs THEN an albatross is a bird' is represented as:

Flies (Albatross) & (Lays eggs (Albatross) → Bird (Albatross)

where → is the implication sign.

Predicate logic distinguishes between situations that are universally true and those that are only sometimes true. The statement 'All feathered creatures are birds' can be expressed as:

$$\forall x [\text{Feathered}(x) \rightarrow \text{Bird}(x)]$$

The inverted A represents ALL, and the formula can be read: 'For ALL entities x, IF x is feathered THEN x is a bird'. On the other hand, the statement 'Some swans are black' is not universal. SOME is expressed using the backward E, thus:

$$\exists x [\text{Swan}(x) \& \text{Black}(x)]$$

The formula can be read: 'There exist one or more entities x, such that x is a swan AND x is black'.

Such formulae can be used to make inferences that will establish the validity of a new formula. For example, suppose that the following knowledge is available to a system:

1. Marcus was a Pompeian: Pompeian (Marcus)
2. All Pompeians died in AD79 (when Vesuvius erupted):
 $\forall x [\text{Pompeian}(x) \rightarrow \text{Died}(x, 79)]$
3. It is now AD1986: Now = 1986
4. All who are dead at a time t are not alive at that time:
 $\forall x \forall t [\text{Dead}(x, t) \rightarrow \sim \text{Alive}(x, t)]$
5. All who died at time t1 are dead at time t2 if t2 > t1:
 $\forall x \forall t1 \forall t2 [\text{Died}(x, t1) \& (t2 > t1) \rightarrow \text{Dead}(x, t2)]$
6. 1986 > 79

The question is asked: 'Is Marcus alive now?' An answer can be derived by reasoning from the above facts. Start with the statement 'Marcus is *not* alive now', and try to infer it. If: $\sim \text{Alive}(\text{Marcus}, \text{now})$, then by substituting in fact 4 we infer: $\text{Dead}(\text{Marcus}, \text{now})$. If this is true, then by fact 5 we infer: $\text{Died}(\text{Marcus}, t1)$ where $t1 < \text{now}$. We infer from facts 1 and 2 that: $\text{Died}(\text{Marcus}, 79)$. We know from fact 3 that now = 1986, and from fact 6 that $79 < 1986$. So it is true that: $\text{Died}(\text{Marcus}, t1)$ where $t1 < \text{now}$. Working back up the chain, this implies the truth of: $\text{Dead}(\text{Marcus}, \text{now})$, and hence the truth of: $\sim \text{Alive}(\text{Marcus}, \text{now})$. Our answer to the question is 'no'.

Representation of knowledge by predicate logic formalism can thus be used in question-answering systems.

FROM PREDICATE LOGIC VIA FRAMES TO SEMANTIC NETS

It has been noted earlier that the same logical statement can be represented in different ways: an IF, THEN statement can be expressed as a predicate logic implication. A further example of transforming a knowledge representation is given here, taken from Nilsson.¹⁶

The following sentences are shown in possible predicate logic form:

1. John gave Mary the book: Give John, Mary, Book).
2. John is a programmer: Occupation (John, Programmer).
3. Mary is a lawyer: Occupation (Mary, Lawyer).
4. John's address is London: Address (John, London).

If these statements were part of a database, it might be convenient to rearrange them so as to group information about John into a unit, thus:

John
 Give John, Mary, Book)

Occupation (John, Programmer)
Address (John, London)
and similarly for Mary.

We could make a further transformation, using the concepts of 'giving-events', 'occupation-events' and 'address-events', each of which is a set with elements. One result would be:

G1
EL (G1, Giving-events)
EQ [giver (G1), John]
EQ [recip (G1), Mary]
EQ [obj (G1), Book]

This expresses the knowledge that statement G1 is an element (EL) of the set 'Giving-events' and that it is equivalent (EQ) to three components: a giver (John), a recipient (Mary) and an object (Book). These component functions are clearly analogous to the participant roles considered earlier. The remaining results of the transformation are:

OC1
EL (OC1, Occupation-events)
EQ [worker (OC1), John]
EQ [profession (OC1), Programmer]

OC2
EL (OC2, Occupation-events)
EQ [worker (OC2), Mary]
EQ [profession (OC2), Lawyer]

ADR1
EL (ADR1, Address-events)
EQ [person (ADR1), John]
EQ [location (ADR1), London]

An alternative form of knowledge representation regards each unit as a 'frame' containing slot names and values (Minsky). The representation above readily transforms into this mode:

G1 (name of frame)
EL of: Giving-events
giver: John
recip: Mary
obj: book

and similarly for the other units.

One further transformation is shown in Figure 1, which is an example of the knowledge representation known as a semantic net (Nilsson). A few other entities have been added (Jobs, Persons, Physical objects). The encircled nodes represent the slot values, and the labels on the connecting arrows (arcs) are the slot names.

Semantic nets may be used to establish links between nodes. For example, the question 'What did John give Mary?' may be answered by following up the 'giver' arc from JOHN, and the 'recip' arc from MARY to reach G1 in each case, and then following the 'obj' arc down from G1 to BOOK.

Another use of semantic nets is to interrelate different classes of entity, their members and their attributes. Consider the example in Figure 2.

ROVER and FIDO are elements of (members of) the class DOG, and CLYDE of ELEPHANT. Both DOG and ELEPHANT 'isa' (is a subclass of) MAMMAL. We see that normally (by default) a mammal is furry and has 4 legs. If we ask the question 'How many legs has Fido?' we follow up the arcs to find 4. But the same question about Rover immediately encounters an exception node - this dog has

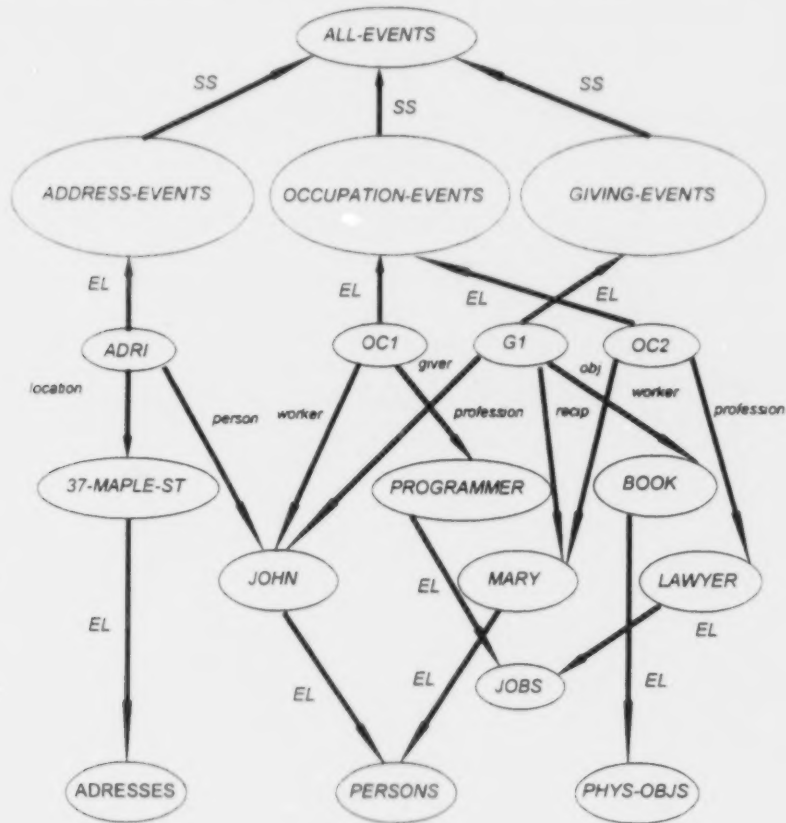


FIGURE 1 A semantic net (from Nilsson, 16 p. 371)

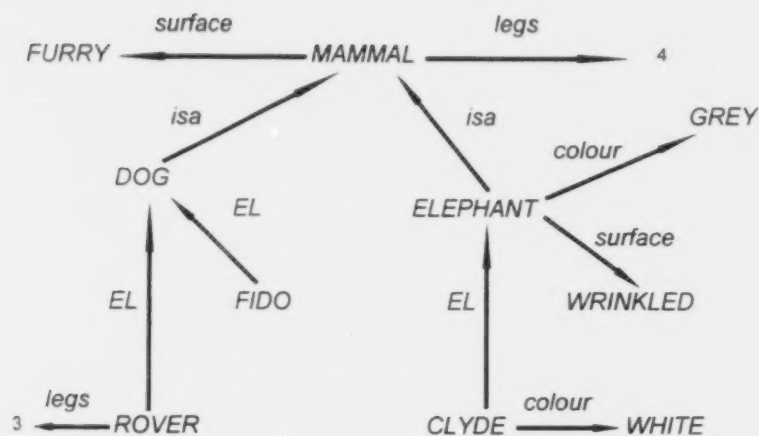


FIGURE 2. A semantic net representing class members and their attributes

only 3 legs. Similarly, though an elephant is usually grey, Clyde is an albino. Both Rover and Fido inherit the default property of being furry, but for elephants (including Clyde) there is the exception node WRINKLED.

RETRIEVAL USING CATEGORIES, CLASSES, FRAMES AND PRODUCTION RULES

An 'intelligent' computer-based retrieval system may be expected to carry out such tasks as

- (a) amplifying a formal query initially received, and transforming it into a search statement
- (b) searching a database
- (c) if the search is unsuccessful, reformulating the search statement

Around these central tasks are others (such as transforming natural language input into a formal query, disambiguating query terms that have two or more meanings, and interacting with the user to evaluate the search), but it is upon tasks (a) and (c) that we will concentrate here.

A prototype expert system for reference and referral has been developed by Vickery, Brooks and Robinson,¹⁷ that makes use of several of the forms of knowledge representation discussed above. The system achieves an 'understanding' of a user query by assigning each input term to one of about two dozen semantic categories. Knowledge of dependency relations between categories guides the system both in attempting to amplify the initial query and in dropping terms during reformulation of the search statement. A classification of terms permits the system to move within an 'isa' hierarchy to broaden or narrow a search. The mechanisms used to carry out these procedures are frame structures and production rules.

The system contains a dictionary of about 1,000 terms likely to occur in queries about its subject domain. With each term is recorded its semantic category, and its class number in a hierarchical classification. Eleven basic categories are used, as follows:

- Objects
- Parts (of an object)
- Processes (behaviour of an object)
- Interactions (between objects)
- Operations (by humans on objects)
- Instruments used in operations)
- Environments (of objects)
- Attributes (of any of the above)
- Uses (of objects)
- Time
- Location (geographical)

Within the present subject domain of the system (gardening), these have been refined into twenty-eight sub-categories, thus:

OBJECTS	plants
	animals (includes insects)
	micro-organisms
	plant diseases and symptoms
	general plant groups
	weeds
	soils
	biological concepts

	miscellaneous
PARTS	of a plant
PROCESSES	occurring in plants
	affecting part of plant
INTERACTIONS	between plant and other object or environment
OPERATIONS	on plants
	on pests and diseases
	on environment
	miscellaneous
INSTRUMENTS	used in any of these operations
ENVIRONMENTS	natural
	man-made
ATTRIBUTES	of plant, their parts, plant groups and collective plant entities
	of processes, interactions and operations
	of instruments
	of environments
USES	of plants
TIME	
LOCATION	(geographical)

Each term in the initial user statement is looked up in the dictionary, and its category and class number noted. This data is now entered into the slots of frames which essentially embody all the possible elements of the facet structure of the subject domain. The system now attempts to fill in further slots in the frames by questioning the user. To do this it is guided by a series of production rules, of which the following is an example:

IF the slot for OPERATION on plant is occupied
 AND the slot for OBJECT (plant) is empty
 THEN ask the user whether a plant name can be supplied .

Further rules give criteria for halting this process - there is no attempt to complete the whole of each frame. When these criteria are met, the system extracts the data from the frames, creates a Boolean search statement from the terms, and searches a computer-indexed database. If the search is not successful, the search state ment is modified: (a), by dropping terms - in a sequence that is in part guided by the category dependency relations, and (b), by moving within the classification hierarchy to extract broader or narrower terms.

This account of the expert system is very much simplified, aiming only to indicate how different types of knowledge representation have been integrated to tackle the practical task of 'intelligent' retrieval. There are a number of other representations discussed earlier - for example, predicate logic, semantic primitives, property inheritance in semantic nets - that also have a role to play in information science.

REFERENCES

1. HURFORD, J.R. and HEASLEY, B. *Semantics*. Cambridge: Cambridge University Press, 1983.
2. FILLMORE, C. The case for case. In: BACH, E. and HARMS, R.T., eds. *Universals in linguistic theory*. New York, London: Holt, Rinehart and Winston, 1968.
3. LINDSAY, P.H. and NORMAN, D.A. *Human information processing*. 2nd ed. New York: Academic Press, 1977.
4. COSTELLO, J.C. and WALL, E. *Recent improvements in techniques for storing and retrieving information*. Wilmington: du Pont de Nemours Inc., 1959.
5. VICKERY, B.C. *Classification and indexing in science*. 3rd ed. London: Butterworth, 1975.
6. FARRADANE, J.E.L. A Scientific theory of classification and indexing. *Journal of Documentation*, 8, 1952. 73-92.

7. VICKERY, B.C. *On retrieval system theory*. 2nd ed. London: Butterworth, 1965.
8. DE GROlier, E. *A study of general categories*. Paris: UNESCO, 1962.
9. WILLETTS, M. Investigation of the nature of the relations between terms in theauri. *Journal of Documentation*, 31, 1975, 158-184.
10. PERRY, J.W. and KENT, A. *Tools for machine literature searching*. Interscience, 1958.
11. SCHANK, R.C. *Conceptual information processing*. Amsterdam, Oxford: North Holland, 1975.
12. WILKS, Y. Good and bad arguments about semantic primitives. *Communication and Cognition*, 10, 1977, 181-221.
13. BUCHANAN, B.G. and SHORTLIFFE, E.H., eds. *Rule-based expert systems*. London: Addison-Wesley, 1984.
14. WINSTON, P.H. *Artificial Intelligence*. 2nd ed. London: Addison-Wesley, 1984.
15. RICH, E. *Artificial intelligence*. New York: McCraw-Hill, 1983.
16. NILSSON, N.J. *Principles of artificial intelligence*. New York: Springer, 1982.
17. VICKERY, A. et al. *Expert system for referral, final report of the first phase of the project*. London: University of London, Central Information Service, 1985.

VICKERY: KNOWLEDGE... This paper first appeared in *Journal of Documentation* Vol. 42, No. 2, 1986 p. 145-159 (ISSN 0022 0418) and has been reproduced here with the kind permission of the original publisher, ASLIB, the Association for Information Management, Information House, 20-24 Old Street, London EC1V 9AP, United Kingdom.

NATURAL LANGUAGE PROCESSING FOR INFORMATION RETRIEVAL

David D. Lewis, Karen Sparck Jones

AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, N.J. 0797; email: lewis@research.att.com
Computer Laboratory, University of Cambridge, New Museums Site, Penbroke Street,
Cambridge CB2 3QG, United Kingdom; email: sparck-jones@cl.cam.ac.uk

The need for automatic *text retrieval* (TR), also known as *document retrieval* (DR) has caught the attention of researchers in natural language processing (NLP). This article explores DR's key properties, summarizes past experience in the field, and reviews various specific NLP research strategies targeting this form of information processing.

For the purposes of this article, we generally view as synonymous the older term DR and the newer term TR. Both involve retrieving texts - from paragraph to book length - for humans to read. In the past, DR pointed the reader to an offline document, usually a journal article or report. New technology makes it practical to store, search, and retrieve online all or part of a document's full text. However, because the technical requirements of the two are the same, we view DR as the general term, except when the distinction just made is relevant and when it is necessary to refer to TR as supplying the user directly with "end" text. We rise information retrieval (IR) as a global term covering everything from DR itself to knowledge retrieval. This article concentrates on IR and on DR as an NLP task.

DOCUMENT RETRIEVAL

Within IR, DR is an important and proper task with its own distinctive properties, not to be confused with data or knowledge retrieval. For example, DR is for the user who wants to learn something by reading about it, as opposed to merely wanting a specific data item or question answered. Just because a user wants to read about, for example

cheap production methods for simple prefabricated housing

does not mean the user has specific questions in mind, like

What are cheap production methods ...

or

How do cheap and expensive methods ... differ?

Moreover, even if the user does have questions in mind, the aim is to retrieve overall information so these questions, as well as other questions prompted by reading the documents, are answered. This means that DR must find relationships between the information needs of users and the information in the documents - both considered in a general sense and neither directly available to the system.

Equally important, the relationship between the user's need and the text that meets it is not necessarily obvious. For example, users looking for information on production methods for prefabricated housing may find answers in

J. Kirk: Reed Mat Huts of Madagascar: Design and Construction.

Retrieval depends on *indexing*, that is on some means of indicating what documents are about. Indexing requires an *indexing language* with a *term* vocabulary and a method for constructing requests and document *descriptions*. Indexing is the basis for retrieving documents relevant to the user's need. It has to be supported by a search apparatus specifying conditions for a match between request and document descriptions, as well as modulation methods, to alter these descriptions if no match is made immediately.

The aim of indexing is to increase *precision*, the proportion of retrieved documents that are relevant, and to increase *recall*, the proportion of relevant documents that are retrieved. In doing so, however, it has to deal with two kinds of problems:

- Those posed by the external context in which the search is done, typically involving few relevant document and many nonrelevant documents; and
- Those imposed by the internal constraints of the task itself, produce uncertainty the retrieval system must overcome.

The main internal constraint is the *variability* in the ways a concept can be expressed⁸. This variability is partly a matter of language, such as

prefabricated vs. unit construction

where the notion of prefabrication is the same, and partly one of perspective, such as

prefabricated vs. factory-made

where views differ on how prefabrication is done. Another constraint is *underspecification* due to vague requests, such as

cheap as economical production vs. cheap as low quality,

or incomplete requests, such as

housing vs. temporary housing.

The difference between these last two examples is that in the first case, users may not recognize their own ambiguity and in the second may fail to provide sufficient detail. This characteristic DR problem stems from the user's ignorance prior to reading. Another constraint is the reduction of documents to their descriptions, which loses information through being indirect, such as

building for reed mat hut,

or partial, such as

construction for design and construction

Because full texts of documents are increasingly available online, the degree of reduction depends increasingly on the indexing method. But reduction is never completely avoided; a document's author always leaves much unsaid on a subject, and reduction is not always pernicious. Compact descriptions of a document's significant content may increase the efficiency of matching and the effectiveness of classifying textual material as relevant or nonrelevant just as feature selection is critical in other classification tasks.

DR thus imposes conflicting demands on text descriptions, asking that they be normalizing, discriminating, and summarizing, as well as accurate. As a result, variations in indexing that increase precision usually decrease recall, and vice versa. Beating this tradeoff by increasing both recall and precision is the fundamental goal of index languages.

There are many types of indexing languages. Terms may be the same terms appearing in the text to be indexed (natural language) or may be limited to those from an artificial or controlled language whose design involves many of the concerns associated with treating meaning representation for NLP.¹ Indexing languages vary according to several factors:

- The form of and emphasis on terms and term relations;
- Whether relations are implicit or explicit; and
- The mixture of *syntagmatic* (document or request specific) and *paradigmatic* (universally asserted) relations.

Natural languages may be used more widely, but hybrids are common, including natural terms combined with artificial relations, such as:

hut MATERIAL (mat MATERIAL reed) LOCATION madagascar.

Wholly controlled forms might appear like this:

(UNIT CONSTRUCTION HOUSING) (MADAGASCAR).

PAST RESEARCH

Tests of a wide range of indexing languages during the past three decades have produced fairly consistent (if not wholly expected) results^{22,24}. These tests have shown that indexing documents by individual terms corresponding to words or word stems produces results at least as good as those produced when indexing by controlled vocabularies, whether simple or complex, and whether produced by manual effort or automatic language processing. Furthermore, automatically combining single indexing terms into multiword indexing phrases or more complex structures has yielded only small and inconsistent improvements over the simple use of multiple terms in a query.

In contrast, statistical DR methods, which ease and enhance use of representations based on single terms, have provided significant improvements over such alternative approaches as Boolean querying¹⁷. Statistical DR methods rank documents based on their similarity to the query or on an estimate of the probability of their relevance to the query, where both query and document are treated as collections of numerically weighted terms. The query can be an arbitrary textual statement of the user's information need or might even be a sample document.

Statistical DR methods assign higher numeric weights to terms showing evidence of being good content indicators, causing them to have greater influence on the ranking of documents. The number of occurrences of a term in a document, in the query, and in the set of documents as a whole, may all be taken into account when computing the influence the term should have on a document's score. In addition, if the user indicates that certain retrieved documents are relevant, this information can be used to reweight and alter the query terms through a process called relevance feedback^{18,21}.

The focus of this baseline statistical DR strategy is on tuning the representation to the current user request, rather than on anticipating future user requests in the document descriptions. This strategy has three major benefits:

- It allows for *late binding*. Complex concepts need not be anticipated during document indexing, but are under the control of each user at query time.
- *Redundancy* is supported by drawing indexing terms from the document text, rather than using a limited vocabulary that may not support a particular user's needs.
- The representation is derived from the documents themselves so that differences and similarities between document texts are given the best chance to survive into the document representations.

For example, a query presented to a statistical DR system might say

A cheap [20] method [5] for prefabricated [30] housing [20].

The term weights 25, 5, 30, and 20 would be assigned automatically to the stems based on their statistics of occurrence in the documents. A document matching the query on the stems *cheap* and *prefabricat* would receive a high score. If the user indicated to the system that this document is relevant, relevance feedback would increase the weights on the words *cheap* and *prefabricat*. In addition, highly weighted terms, such as *unit* and *construct*, from the relevant document might be added to the request with their own weights. They could then promote a hitherto uninspected document through a joint match on *prefabric* and *unit*.

Research showing the effectiveness of statistical DR methods appears solid in tests done in various environments, including different subject domains, ranges of system parameters, say for weighting, and alternative evaluation procedures with distinct performance measures. The methods also generally apply to document routing against standing, rather than one-off needs and perhaps for coarser document categories. However, these studies have involved few documents (30,000 at most, usually fewer) compared to hundreds of thousands in operating DR systems and have largely neglected non-European languages.

Moreover, these experiments have generally been surrogate-based, that is they use titles and abstracts distilled from full-document content with a high loading for what is especially important in the source. The approach also depends on users entering requests in the form of sensible topic specifications while including several terms for alternative matches.

In addition to these caveats to the success of statistical DR, the question also remains of why intuitively plausible improvements in document representation have had so little impact on effectiveness. Why is there so little gain from linguistic sophistication (e.g., from the use of syntactic role relations between terms)? Is it that NLP intended to produce sophisticated indexing has been inadequately done? Is it that our transformations of natural language, even when done well by humans, have been misdirected? Or is it that so much leverage was gotten by searching surrogates at previous experiments that little room for improvement was left? Still, with typical effectiveness results in the range of 30% to 60% recall or precision¹⁷, there is considerable room for improvement, even if DR is an intrinsically coarse process. Further, the research results just described must be considered in the context of operational practice and of the new TR situation where full source tests, and not just their surrogates, are available for direct searching.

THE STATE OF DR

Users can now access thousands of bibliographic databases, mainly in surrogate form, through various services. The long-running debate on controlled vs. natural language indexing has become less important as many commercial databases now use both. Most searches in these databases are done for end users by professional intermediaries who know about database coverage, as well as about the controlled language and indexing practices with it. These intermediaries generally believe a controlled language is superior to natural language, although the controlled languages involve many different design options with no clear winners.

However, searching well-cared-for bibliographic databases is no longer the only function DR deals with. DR sessions can now involve PC users scanning their hard disks for missing files or students searching thousands of Internet servers for an archived Usenet posting. End-user, natural-language searching is inevitable because neither opportunities nor resources are available for using intermediaries and indexers; when full text is available, it seems natural to search it directly.

IR research has been brought to bear against this flood of traditional and nontraditional data, with some

success. Statistical TR system suggested by DR research now range from PCS to 100-gigabyte service databases. However, the situation is far from satisfactory, with at least three classes of problems:

- Uneven penetration of the best methods into operational practice. Many systems still require Boolean logic or other user-befuddling query syntax. When natural-language querying is available, weighting may be unavailable or poorly applied, and relevance feedback is rarely supported. Word stemming operations may also be unsatisfactory or poorly understood.
- Ignorance about applying statistical DR methods to large, heterogeneous databases, particularly to full-text documents. Test collections of this sort only recently became available. Experiments with them have verified that standard techniques work, but many surprises and problems have been encountered^{10,11}.
- Most important, many end users have little skill or limited experience in formulating initial search requests or modifying their requests after observing failure. Even when relevance feedback is available, it needs to be leveraged from a sensible starting point.

Although research shows that natural-language indexing and searching are effective to a degree, users may wonder whether the simple strategies described earlier in this article can be improved. More discriminating methods may be necessary to pluck relevant documents from very large databases and to support the fine-grain definitions of relevance possible with detailed full-text documents. There are therefore two issues. One is whether natural-language indexing, perhaps of a more refined kind than statistically controlled use of single-word terms, is wanted, or whether controlled-language indexing is really what is needed. Both controlled-language indexing and more sophisticated natural-language indexing imply nontrivial NLP, so the other issue is whether the required NLP capabilities are available or in prospect, since large-scale human full-text processing is not a practical proposition.

These issues will be addressed in the context of NLP research, which is itself in an exciting and rapidly changing state. An increase of interest in robust processing, in processing large amounts of real-world text, and in statistical methods in NLP make this an opportune time to consider interactions between DR, and more specifically TR, and NLP.

A TR RESEARCH AGENDA

All the evidence suggests that for end-user searching, the indexing language should be natural language, rather than controlled language oriented. Indexing, or selective text content characterization, is needed, but should be derived from the text, with redundancy and late binding to compensate for uncertainty. For interactive searching, the indexing language should be directly accessible by the user for request formulation; users should not be required to express their needs in a heavily controlled and highly artificial language. This does not mean that the system cannot enhance users' indication of what they want; for example, with statistical data or concept definitions they may not be able to interpret in detail.

Evidence also suggests that combining single terms into *compound terms* representing relatively fixed complex concepts, may be useful. Many controlled languages allow this, and it has been found effective to a degree when done "statistically" on a simple collocation basis in a text window. While compounds uncovered by grammatical analysis have typically been less effective than those found by statistical means⁹, this may change in a TR context. In any case, grammatical and statistical methods are increasingly combined.

The proposal described in the following sections develops these themes and investigates the role NLP may now play in full-text searching. The proposal addresses three things:

- The "words," "phrases," and "sentences" that form individual document descriptions and express the combinatory, syntagmatic relations between single terms captured by the system's NLP-based text-processing apparatus;

- The "classificatory" structure over the document file-as a whole that indicates the paradigmatic relations between terms and allows controlled term substitution in NLP-based indexing and searching; and
- The system's NLP-based mechanisms for searching and matching.

Indexing descriptions. How should we define the linguistic units of indexing descriptions, including the size and depth of text forms and representation forms? For example, should we go for any words or for only nominal group heads, or for concatenated or case-labeled phrases? We propose well-founded simplicity for both the natural-language units taken from the text as inputs to the indexing process and for the natural-language or near-natural-language units in the indexing language descriptions output by the indexing process. Indexing units would be linguistically solid compounds, such as

prefabricated housing

or basic propositions, such as

produce (factory, house).

The success of this proposal depends on its details, which differ from what might be assumed from traditional NLP practice:

- Given the proven value of statistical weighting, any units that NLP produces should be filtered and weighted by the statistics of their occurrences in the database searched and perhaps in other textbases as well^{10,11,15}. Weighting for phrases may differ from weighting for single-word terms to allow for their lower frequency and different distribution characteristics; it is also less well understood than for words^{1,5,6}.
- We also stress the importance of late binding and sensitivity to the uncertainty of evidence. Compound terms will not be identified as definitely occurring or not occurring in a document. Rather, each document will provide some evidence for the presence of each known concept. An occurrence of the syntactically checked noun phrase *prefabricated units* in a document would be good evidence for the presence of the corresponding concept. An occurrence of the verb phrase *(they) prefabricated units* in a document would provide only slightly less evidence for the noun-phrase concept. The occurrence of the two words in separate paragraphs would provide much less evidence, but more than the amount given by the presence of just one of the words or of a related word⁵.
- Basic compound units, such as those just described, would not typically be further combined into frames, templates, or other structured units unless knowledge retrieval (discussed later) is to be supported. The description of a document would be an unordered set of phrases and individual words. This approach applies whether compound terms are formed at document file time or introduced through requests at search time. The rationale is that more complex structures are labor-intensive to design and difficult to fill accurately, and that matches on even basic propositions are so unusual that finer-grained distinctions are unlikely to provide additional information.

Applying the appropriate natural-language procedure to extract all instances of compound terms should produce a reasonable representation for moderately sized documents. For very large full-text documents, further reduction may be needed to produce a summary representation of content not swamped by the idiosyncrasies of numerous subparts. One could restrict terms to those drawn from particular portions of the text or (better) account for both global and local structures of the document when matching¹⁹. Either way, statistical control in unit choice and weighting is required. Only experiment can show what forms of reduction are useful and not too costly.

Thus for processing individual texts, we propose representations in which words and compound terms can refer to concepts with a range of complexity, while the loose coupling among these items permits efficient and flexible matching. Experiments are needed to determine the precise form of these compound terms and how they should be selected and weighted, say, relative to their constituents. NLP

can at least help justify compound-term selection by referring to the grammatical structure of text and perhaps characterizing internal term structure.

Resources used during indexing and searching. Increased recall and precision depend on finding a way for non-identical terms to match. The traditional approach is through *normalization*, replacing several forms with a single canonical form. For example, stemming is a normalization based on morphology:

prefabricated, prefabrication → *pre-fabricat.*

Semantic normalizations are also possible based on manually defined classes:

house, apartment, hut → *DWELLING.*

or automatically detected but previously unrecognized, statistical associations in a document file:

house, lawn, gasoline → *CLUSTER-1738.*

Any normalization applicable to indexing can also be used more flexibly during matching. Retaining original document descriptions has important advantages - notably fidelity - and relational knowledge can be invoked in a context-sensitive and adaptive way during searching. Relationships can be adjusted to suit individual queries either directly (e.g., through user browsing in a graphical display of associations) or indirectly (e.g., through inference from the user's relevance judgements). This strategy also avoids costly reindexing of the entire document file when alterations or additions are made to the system's paradigmatic knowledge.

NLP might also provide various paradigmatic information. In addition, under a model where term relationships suggest, rather than demand, normalization, any resource specifying relationships among terms can also provide paradigmatic information. For example, symbols in knowledge bases, expert-system rule bases, data dictionaries, and source code are usually given names that are natural-language words or compounds. The relationships between terms implied by these structures may be more useful for retrieving text in a particular domain than a general thesaurus.

Procedures for searching. For searching, what mechanisms should be used to set matching conditions and determine request modification? Should matching be loose or tight? Should modulation be free or constrained? Natural simplicity seems to be best, allowing straightforward element stripping or substitution in compound terms, such as using

cheap prefabricated housing

for

prefabricated building

Permitting obvious relational relaxation or substitution is also appropriate, such as using

cause (building)

for

produce (factory, house).

The assumption is that statistics will be applied as a guide or control in iterative searching through selection and weighting. Explicit probabilistic models may be favored over alternative matching schemes for their ability to combine a variety of evidence, but all current models have difficulty dealing with complex descriptions and their elements, so more work is needed.

Comments. Drawing on past DR lessons, we propose that future TR and DR systems include simple flexible natural-language Indexing forms, support devices, and use strategies. Such approaches allow and encourage users to concentrate on request development, which matters much more than document characterization, and to do so in a way that supports derivation, redundancy, and late binding. This approach is also both potentially economically viable, even for large volumes of material, and practical from the user's point of view, given modern interface technology exploiting windows. It is also appropriate for two particular full-text cases:

- Retrieving subtests, including paragraphs: it is still necessary to index on significant concepts even for short, focused pieces of text; and
- Two-level retrieval, first coarse, then fine, allowing motivated zooming.

In principle, many indexing strategies are applicable either at document indexing or request search time, depending on space, speed, and portability factors. For instance, NLP might be restricted to queries, while proximity searching might be used to identify the same compound terms in documents. Besides the efficiency advantages of avoiding natural-language analysis of the document file, an interface applying NLP to queries can enhance access to an existing retrieval system without requiring changes in the system. Another trade-off between efficiency and precision in matching would be to apply NLP only to documents scoring high on a word-based query. Even applying NLP to the whole document file involves trade-offs between explicit indexing on compound terms - speeding queries, but increasing the size of access structures - and indexing only their components or generalizations of their components, such as stems. In other cases, both efficiency and effectiveness may dictate the same course, as when reduction is used in indexing. Careful design of the whole system is required to optimize the related factors, given their interdependencies.

For end users, natural-language indexing strongly related to actual texts is attractive, and while they are required to participate in search development, fast processing and multi-window displays make it easier to exploit available information sources. There are, however, challenges in ensuring that any user understands what is happening and both can and does, for instance, exploit a store of paradigmatic knowledge. It may be difficult to convey the significance of statistical data; and while artificial description forms, like predicate-argument structures, can be applied in TR in a way that is hidden, so users are not confused or repelled, it is still necessary to motivate retrieval output for the user and hence to link the indexing descriptions the system actually uses with comprehensible reports to the user.

IMPLICATIONS FOR NLP

From the NLP point of view, a generic challenge is whether the necessary NLP is even doable; specific challenges are whether nonstatistical and statistical data can be appropriately combined and whether data about individual documents and whole files can be combined, since documents should always be treated in their file contexts.

The demands imposed on NLP by this program differ from those in most NLP tasks. TR, even more than DR, tolerates errors in document representations. In addition, ambiguities in NLP system output, such as alternative compositions of a sentence into phrases, can be assigned probabilities of correctness and used in a probabilistic indexing method⁷. On the other hand, NLP applied to documents must deal with vast amounts of variable-quality text from broad domains. User requests involve smaller amounts of text, but even more variability in form and content. Each of the three main aspects of our strategy - forming text descriptions, providing and exploiting terminological resources, and ensuring matching in searching - poses special NLP challenges.

For example, we left open the issue of which syntagmatic relationships between terms in text would suffice for those terms to form a compound term. Strategies for traditional, if partial, syntactic analysis, allowing processing of hundreds of megabytes of text have been tested for TR^{10,11} but traditional semantic analysis on a large scale has not been demonstrated. New approaches are also possible. Accurate, highly efficient syntactic taggers are available, and some compound terms, like head nouns

and premodifiers, are easily extractable from tagged text³. Various strategies for finding important collocations in large corpora have also been developed²⁰, possibly representing an improvement over traditional IR methods for statistical-phrase formation. Compound terms must not only be generated, but selected and weighted. Methods for exploiting the discourse structure of large texts may be useful in identifying which terms are central to the content of a text.

NLP also has a role in automated and semiautomated acquisition of paradigmatic knowledge. Automated formation of clusters of related words is again attracting the attention of researchers, despite the technique's historical lack of success in DR. More linguistically motivated approaches, such as clustering based on syntactic context, may be an improvement over traditional strategies¹². Leveraging of handcoded resources, such as inducing semantic information from labeled training data or from machine-readable dictionaries, may be a more effective, if less general, approach.

Finally, the type of NLP also constrains the forms of matching that can be performed. For example, element stripping might be restricted to adverbs or to words not in a domain-dependent vocabulary, but these restrictions can be implemented only if NLP has marked compound-term elements with the necessary information. NLP need not be applied identically to queries and documents; the system might do a very careful extraction of compound terms from the request, while the system uses a quick and dirty approach to find compound terms in the vastly larger amount of document text. The resulting uncertainty in the document representation may be compensated for during the matching process. NLP applied to the user request might also help distinguish between request words that should be matched against documents and those that convey other information about user needs, such as *Please retrieve journal articles published after 1987 about...*

A general caution is needed about the prospects that simple NLP strategies will significantly improve IR effectiveness. Recent work in NLP makes heavy use of the context of a word as a clue toward its meaning. For example, methods similar to request/document matching in IR have been used for word-sense disambiguation. It is not surprising that when a document and request match on several words, individual matching words are likely to have the same word sense¹⁴. The matching process itself provides a kind of disambiguation. Another example is that words tend to be accompanied by paradigmatically related words in documents and relevance feedback may add these words to the request, much like a paradigmatic knowledge base would.

Thus NLP techniques are challenged by the basic methods of statistical IR, which has apparently picked some of the low-hanging fruit off the tree. The result is that alternate statistical retrieval methods have had greater impact than alternate text representations¹. This should not discourage research into NLP applications in IR, but suggest that researchers and practitioners examine IR tasks carefully to see where NLP provides added value.

DR vs. DATA RETRIEVAL

Within IR, we distinguish DR from other forms of retrieval. The rest of this article concentrates on the relationships of these other forms to DR, how they may be combined, and how NLP experience may be transferred from one form to another.

We define *data retrieval* as the case where file information is preceded for specific properties and where the conceptual categories for queries must be known in advance.

Natural-language access to databases can replace or supplement the use of formal query languages. It has been investigated for three decades and there are well-established commercial systems⁴. Natural language clearly offers advantages in convenience and flexibility, but also involves major challenges in query interpretation precisely because query expression is decoupled from search formulation. Input queries can require extensive transformation to map onto file categories, and this transformation may have to be mediated by a rich domain model. Ill-formed input can complicate the process. Thus, natural language front ends can be effective, but normally only after significant customization.

The specific difference between DR and data retrieval is that in data retrieval, the set structure for the query is critical and must be specified precisely. The quantificational structure of the input must therefore be identified through natural-language analysis. The user may have a vague query, but the query must still be interpreted in one or more definite ways for searching, like a DR Boolean query. Post hoc set specification, as when the user of a ranked retrieval DR system chooses how far to go down a ranking, is not allowed.

It is not clear that data retrieval experience is directly applicable to DR; the nature of the information base and type of need differ fundamentally, though development of natural-language analyzers for resolving predication structures in data retrieval is relevant to compound-term identification. However, DR techniques that be applied in data retrieval to provide "relaxed" queries automatically if the initial queries do not produce an answer. DR techniques might also be used to generate substitute or "partner" queries for searching accompanying text files. Finally, it is possible that the DR and TR techniques described in this article may be appropriate for databases with Free-text field values and even more for so-called record bases, such as museum catalogues, which can include several free fields (containing up to a paragraph of text), as well as coded or controlled fields.

DR vs. KNOWLEDGE RETRIEVAL

The relationship between DR and knowledge retrieval, or "question-answering," is especially interesting because knowledge retrieval is direct (like data retrieval) but uses less rigorous preceding. Knowledge retrieval thus requires more powerful inference capabilities than either data retrieval or DR.

It is sometimes supposed that replacing a document file with the knowledge base it embodies obviates the need for DR while allowing better IR. This scheme is useful in some contexts, though with Thus NLP techniques are challenged by the basic methods of statistical IR, which has apparently picked some of the low-hanging fruit off the tree. high start-up effort for even limited types of texts, such as banking telexes²⁵. But it is still desirable to access the writer's own presentation, which is one aspect of document content. Presentation is increasingly important for longer texts. and complete replacement with a knowledge-base version is much less feasible.

A potentially more useful strategy is to give DR more depth and integration through an organized superstructure over the file which would be exploited as a knowledge base during initial searching. Document frames, or templates, supported by AI-type inference capabilities would give detailed, consistent, and linkable document characterizations. These structures could be used to regulate query-document matching, guide query modification, and focus browsing. Going further and using a propositional knowledge base would give a unified, high-level collection model, allowing more intensive inference. Many conventional DR approaches, such as faceted indexing and hypertext¹⁶, can be viewed as gestures in this direction. The putative difference would be the explicit automatic inference.

For example, EP-X [16], a search intermediary system using a concept frame hierarchy, is an advance along these lines, but it is based on a controlled language and its knowledge base is constructed manually. Building such bases automatically from documents is very difficult, especially in a way that maximizes the derivation of information from the documents themselves, selects the important information in documents, and manages backup from base to individual documents. Some work has begun in this area^{2,9,12} but in limited domains and by taxing NLP to its limits. Processing is also knowledge heavy, so for wider and larger files, bootstrapping the lexicon is needed.

Although the knowledge base is supposed to encourage query development, which could include question-answering on the base itself, DR suggests that the right approach to knowledge base design is a simple structure embedding natural language, with rich text pointers. The knowledge base would contain semantic structures supporting inferences, but also include pointers to texts:

BUILDING

(TYPE: hut -> text 1

UNIT: mat -> text 1

MATERIAL: reed -> text 1

PLACE: madagascar -> text 1)

Such a structure would be hospitable to user queries and not too constraining. A good case can be made for the same type of structure as a means of linking different bases and base types within global systems. Different bases in such hybrid systems would all be treated as if they were document, or test, collections tied together to support "travels in information space" through associative lexical indexing²³.

CONCLUSIONS

Although conventional DR services continue to make heavy use of strongly controlled indexing languages (like the National Library of Medicine's *Medical Subject Headings*), indexing increasingly involves terms drawn from the natural language of documents.

These simple natural-language indexing techniques have been shown adequate in many experiments, though not on a really large scale. These techniques are also beginning to be used for TR.

However, the greater information detail in full text apparently calls for more sophisticated NLP-based approaches to indexing and retrieval. We suggest that appropriate strategies for this new situation should follow the simple DR methods, extending them to handle compound terms and similar descriptive units. The required NLP technology is being established, and work on applying it to TR is beginning. There are major challenges in making the technology operate efficiently and effectively on the appropriate scale and in conducting the evaluation tests essential to determining whether the approach works and what specific form of it works best²², especially when the tests involve interactive searching of large files. It is particularly necessary to show whether NLP-derived compound terms are significantly better than, say, simple collocational compounds.

The present surge in TR research, stimulated in part by the ARPA-sponsored Text Retrieval Conferences (TRECs)^{10,11}, is a welcome effort. TREC is a major evaluation study with much more data than earlier experiments and comparisons of many different strategies-with and without NLP. However, it is too early to draw conclusions on relative merits, especially since tailoring to the TREC application must be discounted. The retrieval needs covered in the TRECs are by no means typical of many (or most) DR or TR contexts, so care is needed when transferring results, especially since interactive searching is not a primary object of study. Although these tests are on a larger scale than earlier tests, they still involve limitations. More importantly, it is too easy in DR, and hence in TR, to intuit wrongly that things do or will work well, whether these are old approaches, old approaches dressed up in shiny modern technological guises, or truly new approaches. It is essential to test, test, and test again.

REFERENCES

1. Buckley, C. The importance of proper weighting methods. In *ARPA Workshop on Human Language Technology* (March 21-24, Plainsboro, N.J.). Morgan Kaufmann, San Mateo, CA, 1993, 349-352.
2. Chinchor, N., Hirschman, L., and Lewis, D.D. Evaluating message understanding systems: An analysis of the Third Message Understanding Conference (MUC-3), *Computational Linguistics* 19, 3 (1993), 409-449.
3. Church, K.W. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing* (Feb. 9-12, Austin, Tex.). ACL, Morristown, N.J., 1988, 136-143.

4. Copestake, A. and Sparck Jones, K. Natural language interfaces to databases. *The Knowledge Engineering Review* 5, 4 (1990), 225-249.
5. Croft, W.B., Turtle, H.R., and Lewis, D.D. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (Oct. 13-16, Chicago, Ill.). ACM/SIGIR, New York, 1991, 32-45.
6. Fagan, J.L. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*. Ph.D. dissertation, Department of Computer Science, Cornell University, Sept. 1987.
7. Fuhr, N. Models for retrieval with probabilistic indexing. *Inf. Process. Manage.*, 25, 1 (1989), 55-72.
8. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM*, 30, 11 (1987), 964-971.
9. Hahn, U. Topic parsing: Accounting for text macro structures in full-text analysis. *Inf. Process. Manage.*, 26, 1 (1990), 135-170.
10. Harman, D.K., Ed. *Overview of the Third Text Retrieval Conference*. National Institute of Standards and Technology Special Publication 500-225, Gaithersburg, Md., 1995.
11. Harman, D.K., Ed. Special Issue on the Second Text Retrieval Conference. *Inf. Process. Manage.* 31, 3 1995, 269-448.
12. Hindle, D. Noun classification from predicate-argument structures. In 28th Annual Meeting of the Association for Computational Linguistics (June 6-9, Pittsburgh, Pa.), ACL, Morristown, NJ, 1990, 968-275.
13. Jacobs, P.S. and Rau, L.F. SCISOR: Extracting information from Online news, *Commun. ACM*, 33, 11 (1990), 88-97.
14. Krovetz, R. and Croft, W.B. Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10, 2 (1992), 115-141.
15. Lewis, D.D. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (June 21-24, Copenhagen, Denmark). ACM/SIGIR, New York, 1992, 37-50.
16. Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. *Intelligent Databases*. Wiley, New York, 1989.
17. Salton, G. Another look at automatic text-retrieval systems. *Commun. ACM*, 29, 7 (1986), 648-656.
18. Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. *J. American Society for Information Science*, 41, 4 (1990), 288-297.
19. Salton, G. and Buckley, C. Global text matching for Information retrieval. *Science*, 253 (1991), 1012-1015.
20. Smadja, F.A. From n-grams to collocations: An evaluation of Xtract. In 29th Annual Meeting of The Association for Computational Linguistics (June 18-21, Berkeley, Calif.). ACL, Morristown, N.J., 1991, 279-284.
21. Sparck Jones, K. Search term relevance weighting - some recent results. *J. Information Science* 1 (1980), 325-332.
22. Sparck Jones, K. *Information Retrieval Experiment*, Butterworths, London, 1981.
23. Sparck Jones, K. Fashionable trends and feasible strategies in information management. *Inf. Process. Manage.*, 24, 6 (1988), 703-711.
24. Willett, P., Ed. *Document Retrieval Systems* Taylor Graham, London, 1988.
25. Young, S.R. and Hayes, P.J. Automatic classification and summarization of banking telexes. In *Second Conference on Artificial Intelligence Applications* (Dec. 11-13, Miami Beach, Fla.). IEEE Computer Society Press, Los Alamitos, Calif., 1985, 402-408.

About the Authors:

DAVID D. LEWIS is a member of the technical staff at AT&T Bell Laboratories. His research interests include information retrieval, machine learning, NLP, and evaluation. Author's Present Address: AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, N.J. 07974; email: lewis@research.att.com

KAREN SPARCK JONES is Reader in Computers and Information at the Computer Laboratory, University of Cambridge. Her research interests are in natural language and information processing, ranging from linguistic discourse modelling to practical application systems development. Author's Present Address: Computer Laboratory, University of Cambridge, New Museums Site, Pembroke Street, Cambridge CB2 3QG, England; Email: sparckjones@cl.cam.ac.uk

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers, or to distribute to lists requires prior specific permission and/or fee.

LEWIS: This paper was first published in *Communications of the ACM* 39(1) January 1996, p. 92-101 and has been reproduced here with the kind permission of the original publisher.

TESTING...TESTING...

The last paper above ends with the words: "It is essential to test, test and test again." The Dorking Conference gave momentum to the Cranfield experiments whose financial support was being negotiated. The next paper is an account of these experiments by Cyril Cleverdon and Jack Mills, both Dorking delegates. Their work has been debated ever since, particularly perhaps with regard to the definition of relevance. But the concepts of recall and precision and some approximate inverse relationship between the two still provide a rare example of objectivity to our imperfect art. This is evident in the work of TREC in which the spirit of Cranfield lives on.

Alan Gilchrist

THE TESTING OF INDEX LANGUAGE DEVICES

Cyril W. Cleverdon

Director, Aslib Cranfield Research Project and

J. Mills

Deputy Director, Aslib Cranfield Research Project

One day conference, London, 1 February 1963

INTRODUCTION

The evaluation of information retrieval systems has recently become an important matter. In the past, however, most reports or proposals on this type of work appear largely to have ignored the efficiency of operation of the central core of an IR system, namely those operations concerned in the compilation and use of the index. The only aspects to receive consideration are the physical form of the index and the design of thesauri or classifications. The former activity has been slanted towards the use of computers and has tended to assume that this type of equipment will, *ipso facto*, give an improved performance but has made no attempt to justify cost factors which may be one hundred times that of conventional techniques. Work on thesauri and classifications, where it has been practical in nature, appears to consist of compiling lists of terms which go out of favour as quickly as any list of subject headings in the past; the more popular theoretical approach is the setting up of models or the use of increasingly abstruse and complex algebras. From the results and conclusions of the experimental work at Cranfield, it would seem that many of these investigations are comparatively trivial.

In this paper we set out the fundamental operations involved in compiling and using an index, show how the various factors can influence the operating efficiency, and consider the methods to be used in the present Aslib Cranfield investigation.

DEFINITIONS

As the analysis of indexing has become more detailed, there has been an increasing requirement for the more precise definition of the various operations. We have endeavoured to use terms in their conventional meanings wherever possible, but it has frequently been necessary to modify, or to find new terms.

An *information retrieval system* is the complete organization for obtaining, storing and making available information. This could be a definition of a conventional library, but an IR system would be expected to exploit the information in a positive manner and to have extra facilities such as people on the staff capable of evaluating information before it is passed to the inquirer. It would also be expected to have a *subject index* to the items in the store, the index being the physical equipment which permits of the retrieval of references in the searches. The index may be in the form of a card catalogue, a printed list, a set of peek-a-boo cards, a computer, or any other convenient equipment.

The arrangement within the index will depend upon the *index language*¹ and this may be a straightforward alphabetical arrangement of terms, or a classified arrangement of terms, or any variation of these methods. The index language may be used in a *pre-co-ordinate* or *post-co-ordinate* manner. The former implies that the coordination of separate concepts is done at the time of indexing and the entries in the subject index will show this coordination. The latter implies that the coordination of concepts is done at the time of searching, so the entries in the subject index will refer only to single elements.

The *vocabulary* of the index language is the complete collection of sought terms in the natural language, including all necessary synonyms, that are used in the set of documents and are therefore required for entry points to the index language. An *index term*, on the other hand, is an actual term or heading used in the index language, and may be a word or words, as with alphabetical subject indexes, uniterm indexes or zatocoding, or may be notational elements, such as a group of numbers in the Universal Decimal Classification, or may be non-meaningful groups of letters, as in the Western Reserve University Metallurgical Index.

Concept indexing is the intellectual process of deciding which are the concepts in a particular document that are of sufficient importance to be included in the subject index. Conventionally, this involves a 'Yes' or 'No' assessment, for a concept either is, or is not, considered worthy of inclusion in the subject index. It is possible for the indexer to indicate the relative importance of different concepts in a document by *weighted indexing*, which involves the assignment to each concept of a weighting number.

The *exhaustivity* of the concept indexing is a comparative term; at a high level it implies that an entry is made for every possible concept in a document. At a low level it implies that a selection has been made and a smaller number of concepts have been used. *Specificity* is also a comparative term. A concept can be translated into an indexing language in such a way that the index term is co-extensive with a concept. This is a high level of specificity and implies that the index term covers the concept but nothing else besides the concept. Alternatively, the translation can be to a less specific (often called 'broader') index term which includes the concept being indexed as well as other concepts.

Syntactic indexing implies the use of headings which display the relationship between the various elements, as distinct from those which merely show the existence of several attributes relevant to the subject indexed.

A *search programme* is the formalization of the search request and it can show the same characteristics as outlined above for indexing, i.e. it entails a statement of the concepts, which can be at varying levels of exhaustivity, and can be translated into indexing terms of varying specificity.

The *operating efficiency* of an index language will depend upon its performance as regards recall and relevance. *Recall ratio* equals $\frac{100R}{C}$ where C equals the total number of documents in the collection which have an agreed standard of relevance to a given question, while R equals the number of those relevant documents retrieved in a single search. On the other hand, *relevance ratio* equals $\frac{100R}{L}$ where L equals the total number of documents retrieved in a single search.

Operating efficiency is affected by the exhaustivity and specificity of the indexing, as well as the search programme, and by varying any of these factors, one will obtain a performance curve which plots recall ratio against relevance ratio. *Economic efficiency* deals with the performance of the complete index.

¹ In recent papers we have been using the term 'descriptor language' following on the usage of Mr B.C. Vickery. However, Mr Calvin Mooers has pointed out that the word 'descriptor', although now somewhat debased in common usage, originally had a precise meaning. We have agreed to restrict our use of the word to this precise meaning and have therefore decided upon the term 'index language'.

A *set of documents* is any collection of documents which are, or will be, used as the basis of a single subject index. The set can be large or small, restricted to an organization's research papers, or be a heterogeneous collection of journal articles, research reports, patents, etc., in many different languages, but homogeneous in that they will be used as the raw material of a single index. A *set of questions* is a collection of questions to be put to a single subject index, either at present, or at any time in the future when the index is still intended to be operating.

THE PREPARATION OF AN INDEX

Assuming there is agreement concerning the set of documents to be indexed, the following operations have to be carried out in compiling and using an index:

1. Assess the subject matter of each document in relation to the requirements of the users, and decide which subjects should be included in the index. This is concept indexing, and is at present, and for the foreseeable future, an intellectual operation. With a pre-co-ordinate index, it is also necessary to decide on the appropriate combinations of concepts and, if the index language is to show relationships, the syntax.
2. Translate the subject concepts into the index language. This is a clerical task, except in those cases where a new term has to be added to the vocabulary of the index language.
3. Place the indexing decisions into the index, which may involve preparing and filing catalogue cards, punching holes in a card or cards, or making marks on tape. This again is a clerical process.
4. Make a concept analysis of the question and decide on the priority of alternative search programmes. As with concept indexing, this is an intellectual process.
5. Translate the search concepts into the index language, a purely clerical task.
6. Operate the physical retrieval mechanism of the index.

The Aslib Cranfield Project has been primarily concerned with operations 1, 2, 4 and 5, and it has only been due to the necessity of having the index in the physical form that we have been involved with 3 and 6. It is certain that these two latter points play no part in deciding on the operating efficiency, except in so far as that one technique might be more, or less, prone to clerical errors than another. They can, however, significantly affect the economic efficiency of an index.

Involved in these operations is the variable of the index language. Whichever type of index language is used, it is certain that all the stages 1 to 6 have to be carried out. More important is it to note that the only two operations which have a true intellectual content are completely divorced from any consideration of the index language.² The basic concept analysis of the document and the basic concept analysis of the question, with the auxiliary decision of which concepts should be included in the index or the search programme, will be the same irrespective of which index language is used.

It is probably the case that many indexers tend to think in the terms of the index language and their concept indexing decisions may be thereby influenced, but fundamentally it is true that concept indexing is a separate process which should not be affected by the index language.

² It is, of course, true that the compilation and maintenance of the index language can fairly be said to be an intellectual task. Its use, however, within the context of an indexing operation is a separate matter which requires only clerical operations.

INDEX LANGUAGES

The common basic requirement of all index languages is a complete vocabulary of all the sought terms, including all necessary synonyms, that are used in the indexing of a set of documents. This may be likened to an uncontrolled set of uniterms, and must be the basic structure for all index languages; and, whatever ultimate form an index language may take, it can only operate at maximum efficiency by having such a vocabulary. To this basic structure can be added a number of devices which are intended to improve the recall ratio or the relevance ratio. These devices (see Vickery [1]) can be listed as follows:

A. *Devices which, when introduced into an uncontrolled vocabulary of simple terms, tend to broaden the class definition and so increase recall*

1. Confounding of true synonyms.
2. Confounding of neat synonyms; usually terms in the same hierarchy.
3. Confounding of different word forms; usually terms from different categories.
4. Fixed vocabulary; usually takes the form of generic terms, but may use 'metonymy' for example, representing a number of attributes by the thing possessing them.
5. Generic terms.
6. Drawing terms from categories and, within these, facets; this controls the generic level of terms, and to a certain degree controls synonyms.
7. Representing terms by analytical definitions (semantic factors), in which interrelations are conveyed by relational affixes or modulants; the generic level will usually be more specific than when control is by categories.
8. Hierarchical linkage of generic and specific terms, and, possibly, of coordinate terms.
9. Multiple hierarchical linkage, i.e. linking each term to a number of different generic heads.
It should be noted that devices 8 and 3 are not usually (as the others are) methods of class definition determining the structure or constituents of individual subject descriptions; they are ancillary devices (manifested as systematic sequence, or classified arrangement, as a thesaurus, as a network of *see also* references, etc.) indicating the existence of classes wider than these individual descriptions.
10. Bibliographical coupling, and citation indexes; these, also, are ancillary devices which indicate the existence of wider classes, the latter reflecting the use made of the documents and a probability of relevance arising from this.

B. *Devices which tend to narrow the class definition and so increase relevance*

11. Correlation of terms: although implicit in some form in all practical indexing, this device is not inevitable; i.e. the use of a single term to define a class may retrieve quickly and economically, if the term is sufficiently rare in the context of the system.
12. Weighting, i.e. attempts to express the particular relevance of each concept used in indexing a document to the whole document. It may take two forms:
 - i. An attempt to assess subjectively the relative 'information content' of each term within the context of the system;
 - ii. An objective measure, based on statistical counting of the word frequencies, etc.
13. Indicating connections between terms (interlocking):
 - a. Without explicit expression of particular relations (interfixing); this may take at least three forms:
 - i. Partitioning of the document; e.g. if the same document deals with the Conductivity of titanium and the Hardness of copper at a particular temperature, partitioning makes it clear that the document has at least two separate 'themes'.
 - ii. Interfixing within a theme (or 'information item'); e.g. Lead (1) Coating (1) Copper (2) Pipes (2) makes it clear that the subject is the Lead coating of copper pipes and not the Copper coating of lead pipes.
 - iii. If terms are recorded physically in a linear sequence, a citation order (a regulated sequence in which terms from different categories are cited), will convey relations.

- b. With explicit expression of particular relations; this is necessary in cases where simple interfixing cannot cope with the possible ambiguities; e.g. to distinguish which particle is the projectile, which the target, and which the product, in a report in nuclear physics. Two forms are usually recognized:
 - i. Role indicators: these are usually limited in number and express the basic or most common relations found in the subject field concerned; e.g. Product, Starting material.
 - ii. Relational terms: these are more freely developed, the name of a relation being used in the same way as any other term. It may, however, be cited only in the framework of a limited number of fundamental relations, as in Farradane's system.

A third form sometimes to be found in conjunction with *a, iii* above is that of distinctive facet indicators which convey clearly the exact category to which the next term belongs.

Any operating index language is an amalgam of some of these devices; for example a modern faceted classification uses hierarchical linkage within its facets, combines (correlates) terms from different categories (according to a strict citation order so as to maximize the mutual exclusiveness of its classes), provides a degree of multiple hierarchical linkage through its relative alphabetical index, controls synonyms, may control near synonyms, and may confound word forms to a mild degree.

A machine-orientated retrieval system may use correlation as its basic device, but supplement this with confounding of synonyms, confounding of near synonyms, of word forms, show hierarchical and multiple hierarchical linkage to some degree (often via a thesaurus), and use links and roles:

A simple, manual Uniterm system will use correlation as its basic device, but may add to this the confounding of true synonyms, of near synonyms and (possibly) of word forms, a modest degree of hierarchical linkage, and roles.

PERFORMANCE MEASUREMENT

The technique evolved in the Aslib Cranfield tests for the measurement of operating efficiency depends on two factors - the percentage of relevant documents retrieved as against the total of relevant documents in the collection (recall ratio), and the percentage of relevant documents among those actually retrieved (relevance ratio). The original Aslib Cranfield project established the former by conducting searches for questions which had been based on documents known to be in the collection, and the result of this showed that, on an average, about 80 per cent of the source documents were being retrieved. It could be said that this figure did not have particular significance, since it is obviously possible to obtain 100 per cent recall by looking at every document in the collection. Quite apart from going to this extreme, it would have been possible to improve the figure of 80 per cent by 'broadening' the search programme or, as we should say, by making the search programme less specific and/or less exhaustive. The only control on this figure of 80 per cent was that a limit was put on the level of exhaustivity of the search programme. (See Cleverdon [2]), but this was rather crude, and (as discussed in chapter 6 of the same reference), an attempt was made to assess the relevance ratio. This proved to be exceedingly difficult, but it was obvious that it was an essential part of performance measurement, and, in the test of the WRU metallurgical literature [3] the work was done more thoroughly, as is discussed later in this report. The result of a complete analysis of recall and relevance means that it is possible to plot a series of points, depending, for instance, on the search programme, and produce a performance curve as shown in Fig. 1a.

ASSUMPTION FOR FUTURE WORK

A basic premise of the following arguments is that, for a given set of documents and a given set of questions, there will be a maximum retrieval performance; whatever type of index language is used, the performance curves will not be materially altered, and, in fact, the only reason for any major variation from this curve will be inadequacies of the intellectual performance in decisions concerning subject concepts, either in indexing or in searching. For sets of documents in different subject fields, the

resulting performance curves would probably differ, and if the same set of documents were used in two separate situations, since there could be two different sets of questions, this in turn could result in different performance figures. Tests at Cranfield have been concerned mainly with the subject fields of science, engineering, and metallurgy, and with these types of document sets and question sets, it appears that the maximum retrieval performance would be as in Fig. 1a. If the subject matter of the collection were organic chemistry, the resulting performance curve might be as Fig. 1b; for sociology it might be as Fig. 1c.

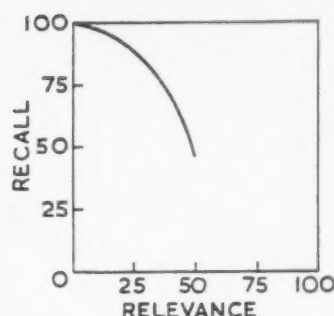


Figure 1a

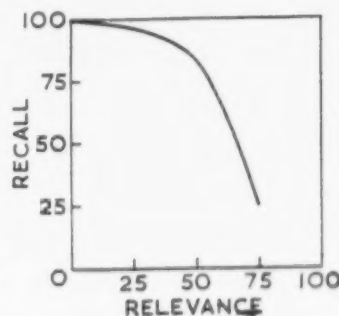


Figure 1b

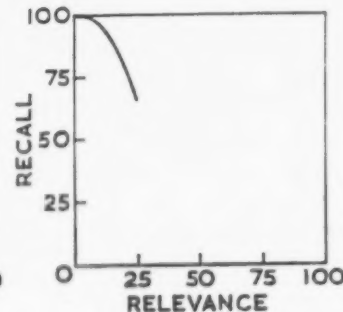


Figure 1c

It has to be assumed in this discussion that we are considering idealized conditions, and do not have to take into account losses due to human error. They have investigated this in the first project, and the general allowance to be made for this is known or, alternatively, can easily be ascertained in any given situation, by using the Cranfield test procedure. We have postulated that there is a fixed maximum performance curve for any given set of documents and questions, and have suggested that, with the type of document and question sets tested at Cranfield, this might range from 100 per cent recall at less than 1 per cent relevance to 10 per cent recall at 30 per cent relevance. The important problem to investigate is how an index can be operated most efficiently at any particular point or within any particular range. As an example, it might be a requirement of an index that it should have a recall level of not less than 95 per cent. Alternatively, the requirement may be that the index should normally operate at a level of 1 per cent relevance ratio, but that it should, when required, be capable of giving a recall figure of 30 per cent.

In such an investigation, there are a number of factors to be considered, and of major importance is the concept indexing. Some groups working in this area seem to place more emphasis on search analysis, in finding how variations in search programmes will affect the efficiency, but this seems to be rather a cart-before-the-horse approach. It is at the stage of concept indexing that the future potential performance of the resulting indexes is determined; if a concept is not included, it will not be possible to recall the particular document by that concept, and vice versa. Concept indexing is, it must be emphasized, an intellectual process that cannot be avoided. Current literature sometimes implies that keyword-in-context title indexing, for instance, is automatic indexing. It is automatic in its preparation of the physical index, but the intellectual stage of concept indexing is still there; the only difference is that in this case the concept indexer is the person who wrote the title.

It is the decisions of concept indexing which result in the level of exhaustivity achieved. The extremes of exhaustivity range from where the whole text is included to where only a single concept is indexed for each document. The effect of high or low exhaustivity on the basic performance curve can be readily shown. 100 per cent recall can always be obtained by doing no indexing but by having available for searching the text of every document in the collection. On the other hand, if only a single concept from each document is included in an index, then it is apparent that there will be a considerable drop in recall. We can, therefore, say that, if all other factors are held constant, the effect of lowering the level of exhaustivity will be to drop the recall figure.

This will in turn result in an improved relevance figure, and briefly it can be said that a high level of exhaustivity of indexing results in high recall and low relevance;

a low level of exhaustivity of indexing results in low recall and high relevance.

Little work has been done on the effect of exhaustivity of indexing, and practice varies widely (see chapter 7 of reference [4]) without the appearance of any positive reasoning behind the decisions. Inevitably a high level of exhaustive indexing results in higher input costs, and this could be a determining factor with certain physical forms of index. In the Western Reserve University test an attempt was made to ascertain the effect of varying the exhaustivity with the facet index that was prepared at Cranfield. In this index the average number of entries originally included for each document was $12\frac{1}{2}$ and the effect of reducing this by stages to an average of three entries per document was investigated. Table I shows the effect that this had on recall and relevance of the documents.

Table I

RECALL AND RELEVANCE RATIO IN FACET CATALOGUE OF WRU TEST
AT VARYING INDEX ENTRIES

No. of entries	No. of documents retrieved	Relevance 2		Relevance 2 and 3	
		Recall ratio	Relevance ratio	Recall ratio	Relevance ratio
		%	%	%	%
12½	891	83	16.3	64.1	34.6
8	824	80.6	17.1	60.6	35.5
5	643	73.1	19.9	50.7	38
3	491	64.6	23	42.1	41.3

This effect will operate irrespective of which index language is used, and, in passing, it should be noted that the same situation prevails in reverse with the decisions concerning concepts to be used in the search programme. However, before going into the matter of search programmes, we will consider the translation of the concept indexing into the index language.

The aspect of the input stage which is concerned with index languages is specificity, and we suggest that, in so far as their operating performance is concerned, a fundamental difference - and probably the most important difference - between index languages is their hospitality for specificity; this we now consider in its simpler form and show its direct relationship to the index language. It is assumed that an index language includes, as a minimum, all the categories of the subject field, whether given in detail or not, of the documents being indexed.

The first stage, i.e. concept indexing, with its decision concerning the level of exhaustivity of indexing, is assumed to have been completed for a given document and we have concepts A_1, B_1, C_1, D_1, E_1 , and F_1 . The next stage is translating these into the index language. Assume three hypothetical index languages, I, II, and III containing respectively, 10,000, 1,000 and 100 index terms. In index language I, there is a straight translation of the concepts into $A_1(I), B_1(I), C_1(I), D_1(I), E_1(I)$, and $F_1(I)$. Index language II translates into $A_1(II), B_1(II), C_1(II), D_1(II), E_1(II), F_1(II)$ (where A is the containing head for A_1, A_2, A_3 , and B the containing head for B_1, B_2 , etc.). Descriptor language III translates into $ABC(III)$ and $DEF(III)$ (where ABC is the containing head for A, B, C, A_1, A_2 , etc.).

The level of specificity which can be given to the indexing is obviously very different in each of these three index languages; the result is that in a search where the requirements are for A_1, B_1, C_1 , although

index language III will retrieve all the relevant documents that are retrieved by index language I, it will also bring out, for example, documents coded A_2 , B_2 , C_2 , resulting in a larger number of irrelevant documents and therefore a lower relevance ratio. In reverse, though the relevance ratio of a search in index language I will be relatively high, it is likely to miss some of the relevant references which would be found by the same search programme by index language III. It might, therefore, be the case that the normal operating level of index languages I, II and III, with all other conditions fixed, would be as shown by points 1, 2 and 3 in Fig. 2. Again, one notes that a lowering of relevance results in higher recall, and of specificity it can be said that

a high level of specificity in the index language results in high relevance and low recall;

a low level of specificity in the index language results in low relevance and high recall.

Considering again the basic performance curve (Fig. 3), we have shown that the maximum recall figure which can be obtained is limited by the level of exhaustive indexing; the maximum relevance figure is limited by the specificity possible in the index language. Therefore, a system which combines a high level of exhaustive indexing with the use of a highly specific index language would have the potentiality, dependent upon the search programme, of operating over the wide range of the curve between points A and B. However, a system which combines a low level of exhaustivity and a less specific index language might not be able to operate beyond the narrower limits shown by points C and D, whatever type of search programme was used. However, if this limited range gives a satisfactory performance figure, then it will be obviously more economical to operate.

It is now possible to consider the effects of exhaustivity and specificity in the search programme. Assume that the concept programme of a question requires that a search should be made for A_1 , B_1 , C_1 and D_1 . If it is decided to 'broaden' the search, then this can be done in two ways. On the one hand, it is possible to drop one of the concepts, so that the search is now for A_1 , B_1 and C_1 . This we would describe as making the search less exhaustive. Alternatively, it is possible to make the search less specific, by substituting A for A_1 , or B for B_1 , (where A is taken to be an inclusive term covering A_1 , A_2 , A_3 , etc.). It is obvious that by broadening the search in either of these ways, the recall ratio will be improved. In particular, however, we should consider the matter of specificity in indexing and specificity in the search programme. To return to the index languages considered earlier, if one has coded according to index language III, then it is useless to attempt a programme that has a high level of specificity. In fact, it is impossible to do this, since the index language does not include index terms of the required specificity.

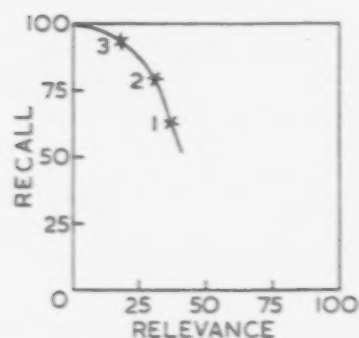


Figure 2

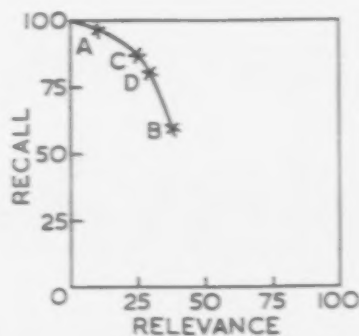


Figure 3

However, with index language I, it is possible to have a highly specific search programme, which would give a high relevance ratio but a low recall ratio (i.e. point I in Fig. 2). If it is desired to improve the recall ratio, then the search programme can be made less specific, until it reaches a level of specificity of the term codes of index language III.

By this time it can be assumed that the recall ratio will have risen to the same level as with index language III (i.e. point 3 in Fig. 2). It is, therefore, shown that, as far as operating efficiency is concerned, there will be for every index language - depending upon its hospitality for specific indexing - a maximum relevance ratio which cannot be exceeded. It will, however, always be possible (by varying the search programme) to improve the recall ratio along the fixed performance curve up to its maximum level, this being dependent on the exhaustivity of the indexing.

AIDS TO EFFICIENCY

If, as is not wholly unreasonable, we consider as a median in the range of indexing languages, one which consists of nothing except an uncontrolled vocabulary of sought terms, then many of the indexing devices mentioned in an earlier section of this paper can be seen to be working in opposing directions. Any device which reduces the number of index terms is working towards improved recall, with the inevitable result that there will be a fall in relevance. Other devices, such as role indicators, are, in effect, increasing the number of index terms and are thereby improving relevance but decreasing recall. These devices have been discussed in an earlier section and the part they will play in the further work will be considered later, but we would now consider the effect of one type of device which is found in many operating index languages.

In the introduction, we said that much of the present theoretical work being done on index languages was trivial. This comment is based on what appear to be logical deductions from the experimental work undertaken at Cranfield. The theoretical work which we have in mind is all concerned with what might be described as the grouping of the terms in the index language. This work is proceeding in many directions: facet classification, dumps, logical associations, word pairs, etc. To describe all this work as trivial shows that we have moved away from a main stream of development in information retrieval, and it may appear difficult to justify the statement. We would first make the point that we say that this work is trivial and do not say that it is useless. This could be translated into more precise figures by suggesting the following. Assume an imperfect index language, covering the field of engineering and metallurgy, which consisted only of an alphabetical list of index terms, and with a normal operating efficiency of 75 per cent recall and 20 per cent relevance. If one gave to this index language the additional device of a grouping of the terms, then we suggest that the recall ratio of 75 per cent might be achieved with a relevance of 21 per cent or, alternatively, that with a relevance of 20 per cent one might have a recall ratio of 78 per cent.

The purpose of any grouping of terms is to lead the searcher to related terms which might have been used by the indexer for relevant information. It is logical to assume that this is a worthy objective and it has been a common assumption that it is also a vitally important objective. Why is it, in fact, of such little importance? The reason is because it is incorrect to assume that the comparison is between an index language without a grouping of terms and an index language with a grouping of terms. The comparison is, in actual fact, between an index language with a pre-arranged grouping of terms and an index language with a grouping of terms that is special and personal each time it is used. In other words, a searcher, whether a librarian or a subject specialist, is capable, to a greater extent than has been generally appreciated, of making a grouping of related terms at the time of the search and in relation to the particular subject of the search. The problem is that a single classification schedule, a single set of cross-references, a single set of thesauri-type headings, or a single set of any other form of grouping, however logically or abstrusely derived, is most unlikely to be of equal value to all the different types of user of an information retrieval system. With a metallurgist, a chemist, a mineralogist, a mining engineer or a design engineer, a single term might well have five different sets of associations, and the individual concerned will be able to decide for himself, at the time of the search, the set of related terms that would be an improvement on a set of pre-arranged groupings. It would, therefore, appear logical that the best method would be to devise a separate set of groupings for each type of user, so as to maximize the chances of improving recall with a minimum drop in relevance. It should be noted that such groupings can be devised independently of the index language, and can therefore be used with any index language.

It is, however, necessary to qualify the above comments, and also to return to the point, raised earlier, concerning the presumed difference in the shape of the performance curve for subject indexes in different subject fields. If we consider pure data retrieval, of the type where it is required to know, for instance, the names of all persons in an organization who joined more than ten years ago, who are graduates and whose salary is less than £1,800 a year, it might be expected that all the relevant information and only relevant information would be retrieved. As one passes from the field of data retrieval to the field of information retrieval, the terminology becomes less precise in its meaning, there is an increasing number of errors due to confusion and the performance of the subject index becomes progressively worse. We have suggested that chemistry, being more precise in its terminology, would permit the compilation of a subject index having an improved performance curve in comparison with an index in the field of engineering and metallurgy of the type tested at Cranfield. On the other hand, many areas in the social sciences have a vague terminology. For example, in the opening pages of this paper, it was considered necessary to give the definitions of many terms that have been used, and it would take little effort to find examples of the same term being used with other meanings, or other terms being used for the same meaning. In such circumstances, it is hardly surprising that retrieval performance is hampered, and it is reasonable to suggest that there would be an increased requirement for a grouping of the index terms. Workers in the field of engineering and metallurgy appear to be sufficiently well acquainted with the terminology of the subject field to make their own grouping of terms, with the result that an index language which lacks a pre-arranged grouping of terms is only marginally inferior to an index language which has such a grouping of terms. We can, therefore, assume that any subject field that has a more precise terminology than engineering will have even less need for a grouping of the index terms.

THE FUTURE WORK OF THE CRANFIELD PROJECT

It might be said of Cranfield I that its main endeavour was that it tried to measure the operating efficiency of complete indexing languages in a simulated real life situation, and also that it used, as its major measure of performance, a relatively crude gauge of relevance based on the retrieval or non-retrieval of a single source document. Cranfield II will differ in both major respects; firstly it will try to measure the impact on recall and relevance of particular indexing devices - the elements which go to make up a complete index language; and secondly, it will try to use decidedly more precise measures of relevance.

The devices that have been identified are listed earlier, but some further observations on them seem to be called for:

- i. Their significance is that they provide different ways of defining classes, i.e. different ways of indicating relations between the concepts which are the subject of a document. Consequently, they differ in the degree to which they produce broader or narrower classes, and thereby assist or hinder specificity and exhaustivity.
- ii. The function of class definition in indexing is that the assignment of a document to a particular class or classes (x, say) may allow us to ignore other classes (not-x), on the assumption that the kind of document we seek, if it is not in class x, cannot be in class not-x. Classes are established which will predictably include certain documents, and as predictably exclude others. This assumption can never be realized fully in practice, for classes in documentary classification can never be quite mutually exclusive. A documentary subject description ABCDE is never more than a statement of what the document is *mainly*, about. ABCDE does not necessarily exclude FG; these may feature in it in some minor way. A report on 'Laminar boundary layer flow' will reveal, in a single casual aside, half-way through the text, that 'Three-dimensional boundary layer' is assumed. If this concept (three-dimensional) is not indexed, it does not mean that the concept really is banished from the subject content of the article. It is this awkward indeterminacy of the literature to be indexed which we seek to circumvent by exhaustive indexing.
- iii. Different class definitions have different focusing or resolving power. If we take the simple correlation of keywords in the title and abstract of a document as a relatively crude device for establishing classes, then other devices, if added to these, either enlarge the scope of the classes

or constrain them further. For example, to confound verbal forms, so that the class designations 'Heat', 'Heated', 'Heating', 'Hot', are extended to designate a single class covering all these different concepts, would clearly broaden the description. This would inevitably increase recall and reduce relevance - but by how much? Or, if we now introduce role indicators into our index language, so that a class designation 'Zinc' becomes several different classes according to the role played by Zinc in the particular subject description, e.g. Zinc (starting material), or Zinc (product), or Zinc (agent), then these classes are clearly narrower, and recall will now decrease and relevance increase. But again, by how much?

- iv. Certain of the devices, such as generic and specific terms linked hierarchically, might appear to perform a function different from that of mere class definition. The difference between the operation of *arrangement* in an index, which is designed to display relations between ostensibly separate classes (the chief function normally associated with classification), and the operation of class definition, is often overlooked. In traditional classification, different principles of division might be applied to produce different classes, and then these classes are arranged, usually in some systematic order. But the two operations, of division to produce classes, and the arrangement of those classes, are quite distinct.

The functions of arrangement are well known. It may show clearly various connections between classes (and this function is itself a looser form of class definition); it may help the programming of searches, by suggesting other approaches; it may be economically efficient in the performance of these functions by reason of physical juxtaposition of entries or documents. Cranfield I appeared to indicate that the differences in the arrangement of classes (full classification, A/Z direct, A/Z indirect - i.e. alphabetico-classed) are only of marginal importance in retrieval. But we need to remember context here: the context for which the findings seemed to be valid was that of indexing, not the physical arrangement of documents (which rules out the entire field of shelf display); of a special field, rather than a general collection, where interconnections between classes are more difficult to display thoroughly and far more difficult for a librarian to carry around in his head; of indexing by experienced and skilled indexers, with equally intelligent searchers.

Is arrangement, then, a device to be tested here? Hierarchical linking is certainly to be tested, but how it is implemented - whether by a classified arrangement in the index itself, or by a thesaurus, or by systematic connectives in an alphabetical catalogue - is probably outside the field of our inquiry just now. But independently of the convenience of different methods of performing the function of hierarchical linkage, the extension of the field of likely search which this device implies may be regarded as a form of class definition, whereby membership of a given class carries with it an explicit or implicit indication of the existence of a wider class containing it. Much the same may be said of semantic factoring; for example, to index 'Tempering' as a kind of process, acting on a metal, making use of heat is another device for extending the field of likely search, of broadening the effective class definition.

- v. It might be argued that an index language is more than the sum of its parts, and that evaluation of the contribution to recall and relevance of isolated devices will not help to evaluate the complete index languages which are the practical concern of workers in information retrieval. But the evaluation of complete languages is clearly incomplete so long as we do not know which features of these languages are making them work as they do. For example, it could be argued that the good performance of the Universal Decimal Classification in Cranfield I owed much to the unusually thorough A/Z index made for it. It has already been observed that a complete language will invariably contain an assortment of different indexing devices, and the investigation should give some definite measure of the separate contributions these devices make to recall and to relevance.

METHOD OF INVESTIGATION

To obtain reliable measurements of the impact on recall and relevance of the different devices requires careful precautions for their isolation. This will be done by operating in the laboratory conditions which are provided by a set of questions and a strictly limited number of documents providing the answers to

be retrieved, so that there can be a 'closed' collection, in which the relevance to each question of each document in the collection is known. The questions and documents are being assembled as follows.

The assumption is that a research paper is written as the result of an investigation which has been undertaken to provide the answer to a question or questions in the mind of its author; also, that if an author cites references these must have something to do with the subject of the paper. In the field of high speed aerodynamics there are generally from five to ten references attached to each research paper. We are writing to four hundred or more authors of research papers (both theoretical and experimental) in this field. Each author is asked to state, as nearly as possible in the form of a search question, the basic problem which was the reason for the investigation. Also, if there were any subsidiary questions which arose in the course of the work, he is requested to enter up to three of these (see Fig. 4, sheet 1). He is then asked to indicate the relevance of the references to each question in the following scale of 1 to 5:

1. References which are a complete answer to the question. Presumably this would only apply for supplementary questions, i.e. (B), (C) or (D) since if they applied to the main question there would have been no necessity for the research to be done.
2. References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.
3. References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.
4. References of minimum interest, for example those that have been included from a historical viewpoint.
5. References of no interest.

**ASLIB CRANFIELD INVESTIGATION INTO
PERFORMANCE OF INDEX LANGUAGES**

Please complete and return to:

C.W. Cleverdon
The College of Aeronautics,
Cranfield, Bletchley,
Bucks., England

AUTHOR J. F. Clarke
TITLE Reaction-resisted Shock Fronts
REFERENCE College of Aeronautics Report No. 150, May 1961

BASIC QUESTION

- (A) Has anyone investigated the role of chemical reactions in determining shock wave structure?
- (B) Papers on acoustic wave propagation in reacting gases.
- (C) Have other types of non-viscous compression waves been investigated analytically or observed experimentally?
- (D) Is there any simple, but realistic, "model" gas which can be used to expedite analysis?

Figure 4 (sheet 1)

LIST OF REFERENCES

Ref. No.

ASSESSMENT
Question

1. Clarke, J. F. Flow of chemically reacting gas mixtures.
College of Aeronautics Report 117.
2. Clarke, J. F. Linearized flow of a dissociating gas.
Jnl. Fluid Mech., 7, 1960.
pp 577-595.
3. Griffith, W. C. And Kenny, A.
Jnl. Fluid Mech., 3, 1957, 0 268.
4. Lighthill, M. J. Survey in Mechanics
Ed. by G. K. Batchelor and R. M. Davies, Cambridge University Press, 1956.
5. Lighthill, M. J. Dynamics of a dissociating gas.
Jnl. Fluid Mech., 2, 1957, pp 1-32.

A	B	C	D
5	1	5	3
5	1	5	5
5	5	1	5
2-3	3	1	5
5	5	5	1

Figure 4 (sheet 2)

The references will now constitute the basic indexing collection. The original papers which cited the references will be discarded, since the correlation between them and the questions they were designed to answer might be undesirably high. Ultimately, it is intended to obtain a collection of some 1,500 documents and some 400 questions, and for each question there will be some documents with a certain stated degree of relevance to them.

However, it is possible, even probable, that for each question some of the other documents in the collection (i.e. those not quoted as references by the author of the original paper) are also relevant in some degree. The only way to establish this reliably is to examine the whole collection in respect of each single question. This will be done by a number of post-graduate students working in the subject field. Any further documents which they think might be relevant to the question will be sent to the author of the original paper, who will assign relevance values to them as he did to the references in his paper.

What will emerge will be a closed system in which we have a set of documents (about 1,500) and a set of questions (about 400), and in which every document has been checked against every question. Therefore we know, for every question, exactly which documents in the collection are relevant, and to what degree. It might work out, say, that for a given question there will be two documents that are really important, four of which are useful, three which are marginal in relevance, with all the rest irrelevant.

Perhaps the most difficult problem in the above procedure is that of giving a measure of the relevance to a question of a particular document. It is probable that a serious element of subjectivity is likely to remain in any measure attempted. Certainly, to attempt too much refinement here is to risk deceiving ourselves; so we are attempting only the few broad distinctions listed above.

Promise of a more objective measurement of relevance may lie in the degree to which a question description matches a document description, and this is being explored. The distinction between 'relevance' and 'pertinence' drawn in some recent papers is certainly a real one. A document may match a question exactly in the matter of conceptual content (which is what we describe in indexing) and yet be of no use (not pertinent) to the requester because it is in a foreign tongue, is too mathematical in treatment, or has been superseded by another paper. Some of these factors can be handled in the index by means of further description, such as the well-known 'form' divisions of book classification and cataloguing. Others are not measurable in any laboratory investigation (e.g. 'I've read this already').

INDEXING

All index languages consist in the first place of a substantial reduction of the language of the original collection of documents themselves to a relatively small and limited index language. This is true even of automatic retrieval systems in which the whole file is scanned, for the machine has built into it a programme which has already rejected much of the language of the text being scanned, for example its articles, conjunctions, etc.

The indexing devices to be measured may be regarded as refinements, of one sort or another, to be imposed on the crude index language (not just vocabulary, but essential relations also) which is obtained by selecting the significant words in the title, summary and text of the document indexed. So our first task is to establish this crude language for each document, and thus for the whole collection. This must provide the raw material, the basic information regarding the document's subject content, to allow us to index the document subsequently with maximum exhaustivity and specificity.

These terms, exhaustivity and specificity, have already been defined, but further explanation of their practical implications seems desirable. Both for indexing and for searching programmes, we regard the limits of broadness and narrowness to be contained within a framework determined by the specificity and exhaustivity of the descriptions used. In indexing, the specificity of a term refers to the degree to which it reflects the precise generic level of the concept it stands for; for instance, to index a 'slender delta wing' specifically is to index it as just this, and nothing less. To index it as a 'delta wing' is to describe a genus to which it belongs, not the particular species itself - and such indexing is not specific. It might be called less specific or more generic - the terms are purely relative.

This is, of course, nothing more than the traditional genus/species relation, which is sometimes thought to be the only acceptable relation in strict classification. Its significance in indexing is that it refers to an inclusion relation which is permanent in character; slender delta wing, for example, is always a kind of delta wing, being included in it, and a document on the one will always be of some relevance to a document on the other.

Exhaustivity in indexing refers to the degree to which we recognize (i.e. include in our subject descriptions) the different concepts in a subject which are not in a genus/species relation - which, in other words, come from different categories or facets. For example, a paper on the 'calculation of the laminar boundary layer on a slender delta wing' is exhaustively indexed only if all three categories present are recognized, i.e. the kind of aircraft structure, the kind of aerodynamic flow, and the kind of activity directed at the subject (calculation as distinct from experimentation, say). It might be said that exhaustivity covers also the recognition of the different elements in a hierarchy - e.g. of 'slender' as well as of 'delta'. But the idea of specificity takes care of this: a specific indexing description automatically recognizes those links in a chain which constitute sought terms in the indexing sense, as this example demonstrates.

Exhaustivity does cover, however, recognition of different co-ordinate terms from one hierarchy; e.g. exhaustive indexing of a document on the conductivity of metals *a*, *b*, *c*, *d* and *e* would require the separate indexing of each metal.

Except for the Whole/Part relation, these non-hierarchical relations are not inclusion relations, and the connections between members of different categories are not so permanent or so predictable as the connections between members of the same hierarchy, e.g. whereas a search for information on 'Slender, delta wings' would automatically find relevant a document on 'Delta wings', it would not necessarily find a document on 'Boundary layer' to be so.

Earlier on, we have stated that greater exhaustivity in indexing makes for higher recall and lower relevance, whilst greater specificity makes for lower recall and greater relevance. This may puzzle some readers when they consider the effect on a single subject description of these factors. For example, to index a report as being about 'slender delta wings' rather than just 'delta wings' narrows the class definition and so reduces the number of documents in the class - i.e. it improves relevance, but reduces recall. If we index a report as being about 'experiments on the boundary layer of a delta wing', rather than just 'boundary layer of a delta wing', we again narrow the class definition (this time by correlation), and so reduce the number of documents in the class; again, we improve relevance, but reduce recall.

It is clear from this that we must distinguish between exhaustive indexing in the sense of making a single subject description truly exhaustive of the subject of the document, and exhaustive indexing in the sense of providing the maximum number of retrieval handles or access points necessary to full retrieval. An indexing system could recognize all the three categories in the above subject by means of separate entries, one for the kind of wing, one for the kind of flow, and one for the kind of intellectual operation. At no point would any one subject description be completely precise, i.e. exhaustive of the categories in the complete subject. But they will collectively exhaust it and provide all the required access points. It is this second sense of the term 'exhaustivity' which we use here. Indeed, it is open to argument whether the first meaning is a practical possibility in indexing. A single subject description of a complete report, to be truly exhaustive, would need to be, at the very least, of the dimensions of a full WRU type abstract. If, however, we recognize that in practice, thorough indexing partitions a document into its 'information items' (or 'themes' as we tend to call them), then exhaustive indexing, in the sense of a single subject description of a theme being a precise and exhaustive description of that theme, may well be an objective.

SPECIFICITY AND EXHAUSTIVITY IN INDEXING AND SEARCHING

Below are two grids, showing the degrees of exhaustivity and specificity possible in the indexing of a document and in the framing or programming of a search question (which is really a form of indexing of the subject implied by the question) to which that document is relevant. The document and question are taken, with slight simplifications, from reference 2).

Document		
"An approximate method of calculating the laminar boundary layer on a delta wing."		
SPECIFICITY (Hierarchical level recognized)	Exhaustivity (number of categories recognized)	
	Aircraft structures	Operation on subject
	Aerofoils	Investigation
	Wings	Theoretical
	Delta Wings	Mathematical
	Slender delta wings	Calculation
	Aerodynamic entity	
	Flow	
	Laminar flow	
	Laminar boundary Layer	

A linear statement of the precise subject omitting implicit terms would be

Slender delta wings - Laminar boundary layer - Calculation

Varying the specificity of indexing - this implies indexing the terms within one or more categories at levels above that constituting 'specific' indexing, e.g. varying the level in the structures category (by omitting the specification 'Slender') would give

Delta wings - Laminar boundary layer - Calculation

or varying the level in the Aerodynamic Entities category would give

Slender delta wings - Laminar flow - Calculation

It should be noted that part of the hierarchy of both 'Structures' and of 'Flow' could be varied in their order of subordination; i.e. they could be altered to: Wings-Slender-Delta, and to Flow-Boundary layer-Laminar. This possibility might add a complication to a pre-co-ordinate system.

Varying the exhaustivity of indexing - this implies the non-recognition of one or more of the categories, e.g. omission in the indexing of the Operations category would give

Slender delta wings - laminar boundary layer

Omission of the Structures category would give

Laminar boundary layer - Calculation

Question

Three-dimensional boundary layer on slender wings

Aircraft structures	Aerodynamic entity
Aerofoils	Flow
Wings	Boundary layer
Slender wings	3-dimensional boundary layer

A linear statement would be

Slender wings - 3-dimensional boundary layer

Varying the specificity in searching is exactly analogous to varying the specificity of indexing, e.g. varying the specificity in the Aerodynamic Entity category would mean a search for the broader subject

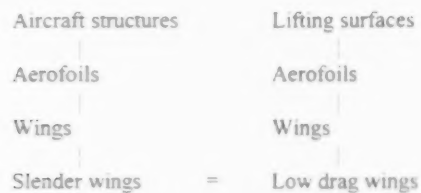
Slender wings - Boundary layer

Varying the exhaustivity in searching is exactly analogous to varying the exhaustivity of indexing, e.g. the omission of the Structures category would mean a search for the broader subject

Three-dimensional boundary layer

The hierarchical relations shown above are incomplete in two ways.

1. *Multiple generic links* are not shown. For example, a slender wing, from another viewpoint, could be characterized as a low drag wing factor. Theoretically any one term may be considered as a species in a number of different hierarchies, and the grid above could be enlarged within each category to show this :



2. *Co-ordinate relations* are not shown, although theoretically these may include potentially useful connections, as *Turbulent* boundary layer on slender delta wing.

In the first place, both the indexing and searching will be exhaustive and specific within the capabilities of a given index language; for instance, our first test might be of a language consisting simply of the correlation of uncontrolled 'uniterms'. In the example above, this would mean a search for the description 'Three-dimensional boundary layer on a slender wing' in a file in which the document is indexed as 'Calculation of laminar boundary layer on a slender delta wing'. We will then vary one element at a time, e.g. we will assume a degree of indexing which is less specific, first within one category, then within two categories, and so on, whilst holding the question exhaustive and specific. At each stage, figures for recall and relevance will be noted. Next, the exhaustivity of indexing is varied, then the specificity of the question, and so on.

After completing this stage, a particular indexing device is introduced which has been absent so far; for example, to the simple correlation of uncontrolled 'uniterms' we may introduce the confounding of verbal forms and note its impact on recall and relevance, again testing within a range of variables of specificity and exhaustivity of indexing and searching.

The number of different permutations of variables implied in the above programme is obviously vast and will constitute one of the major problems. We anticipate, however, that many of the theoretically possible combinations of factors will rule themselves out as having no retrieval or search significance; reducing the specificity of the Structure category to 'Wing', say, and reducing exhaustivity by omitting the Flow category would give us a description 'Wings - Calculation', which is obviously worthless.

PROBLEMS OF BASIC INDEXING

To provide the raw indexing material for subsequent processing we begin by recording the product of the 'concept indexing'. This means that we enter on a master sheet (see Fig. I) a statement of what we consider to be all the significant terms, indicating as far as is necessary the relations between them. This is being entered at four levels of description, namely:

- i. Abstract
- ii. Themes
- iii. Concepts
- iv. Terms

The abstract provides a brief but general statement of the subject content of the complete document. The themes are required to show the relationship of the concepts, and will be used in connection with such indexing devices as partitioning, interlocking and interfiling. With the concepts, it will be noted that the same concept may occur in different themes, and similarly the same term may appear in two or more concepts. It should be emphasized that we have no particular system in mind when doing this initial indexing; it is completely neutral.

B 1069	AUTHOR G. E. Kaartari TITLE Predicted shock envelopes about two types of vehicles at large angles of attack.	Indexer FWL	Date 4/2/63	
Base Document A32	REFERENCE NASA Tech. Note D-860			
ABSTRACT Presents methods for predicting detached shock envelopes of two atmosphere entry vehicles, one a high-drag capsule shape, the other a slender triangular wing (conical vehicle). Predicted and measured shock envelopes were compared for Mach range of 3 to 15 for vehicles at high angle of attack.				
THEMES	CONCEPTS & WEIGHTS	TERMS & WEIGHTS	TERMS & WEIGHTS	
Atmosphere entry vehicle—at high angle of attack—in hypersonic flow (Mach 3 to 15)—detached shock envelope of—theory.	Atmosphere entry vehicle 8	Atmosphere 8	Slender 8	
	High angle of attack 9	Entry 8	Triangular 8	
High-drag capsule entry vehicle—at high angle of attack—in hypersonic flow (Mach 3 to 15)—detached shock envelope of—prediction.	Hypersonic flow 7	Vehicle 8	Wing 8	
	Mach 3 to 15 6	High 9	Conical 8	
	Detached shock envelope 10	Angle of attack 9	Standoff 6	
Slender triangular wing (conical) entry vehicle—at high angle of attack—in hypersonic flow (Mach 3 to 15)—detached shock envelope of—prediction.	Theory 10	Hypersonic 7	Angle 6	
	High-drag capsule 8	Flow 7	Continuity 7	
	Slender triangular wings 8	Mach 3 to 15 6	Mass 7	
Conical entry vehicles—at high angle of attack—in hypersonic flow (Mach 3 to 15)—shock standoff angles—prediction of.	Conical vehicle 8	Detached 10	Wave 10	
	Shock standoff angle 6	Shock 10	Surface 5	
Continuity of mass flow between shock wave and entry vehicle surface.	Mass flow continuity 7	Envelope 10	High 8	
	Entry vehicle surface 5	Theory 10	Shock 6	
		Drag 8	Entry 5	
		Capsule 8	Vehicle 5	

Figure 5

An additional type of statement that we provide is the weighting of the different concepts, and thus the terms. This is a relatively untried device that appears to have the possibility of being a particularly powerful tool in improving the relevance ratio. If one is doing intellectual indexing, it seems reasonable to suggest that the weighted indexing should also be a subjective assessment. If the indexing is to be done by using statistical techniques, then obviously the assignment of weights is a ready by-product.

If the impact on recall and relevance of the various devices is to be isolated, we cannot afford to have variations in other factors introducing uncontrolled influences. We cannot afford a situation, for example, where the figures for the effect of a particular device are criticized as unreliable because the indexer had omitted some vital term or relation. In particular, we want to be sure that all the relevant documents can be retrieved when indexing is exhaustive and specific, i.e. all the facilities indexwise must be available for complete retrieval from the index language (remembering that complete retrieval is always available to any librarian if he is prepared to abandon his index language and search the whole collection). This will be ensured by checking the indexing of a document against the questions to which it is known to be relevant and making quite sure that the raw indexing material does have this necessary data.

It is an implicit assumption of the provisions we intend to make for indexing that the ability of any index language to describe a subject closely or precisely is crucial if a high relevance ratio is required. If the index language (its vocabulary and its syntax) does not allow precise description, then the most exhaustive and specific searching will not improve matters in this respect. It is interesting to note that the age-old arguments about whether a classification should be detailed or not now has the basic issue determined with reasonable precision. While a detailed classification will provide precision for those who need it, but can be used broadly by those who do not, a broad classification provides facilities only for the second eventuality. At the same time, it is possible that the surprising degree of correspondence in the recall performance of the four different systems in Cranfield I can be attributed largely to the care with which the vocabulary was established in each case, even though the syntax of the schemes varied considerably.

It should be noted, however, that the presence or absence of provision for the display of relations will, in certain circumstances, have the same effect on recall and relevance as the presence or absence of provision for vocabulary terms, e.g. a search for *x as an agent* will not be retrieved independently of *x as a patient* unless the index language allows the difference in role to be indicated.

Earlier in this paper, we suggested that the subject area covered by an indexing system might prove to be the most potent factor of all in determining a performance curve. We hope to be able to shed some light on this by testing an index in the field of the social sciences, where we may find that the potency of particular devices will vary significantly from their potency in a relatively unambiguous area like high speed aerodynamics.

REFERENCES

1. **Vickery, B. C.** - (1961).
On retrieval system theory / B. C. Vickery. - London, Butterworths, 1961
2. **Cleverdon, Cyril W.** - (1962)
Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems / C. W. Cleverdon. - Cranfield, [College of Aeronautics], October 1962. Aslib Cranfield Research Project.)
3. **Aitchison, Jean and Cleverdon, Cyril W.** - (to be published).
Report of a test on the index of metallurgical literature of Western Reserve University / Jean Aitchison and Cyril W. Cleverdon. - To be published.
4. **Cleverdon, Cyril W.** - (1960).
Report on the first stage of an investigation into the comparative efficiency of indexing systems / Cyril W. Cleverdon. - Cranfield, College of Aeronautics, September 1960. (Aslib Cranfield Research Project.)

CLEVERDON: This paper first appeared in *ASLIB Proceedings* 15(4) April 1963 and has been reproduced here with kind permission of ASLIB, the Association for Information Management, Information House, 20-24 Old Street, London EC1V 9AP, United Kingdom.

INDEXING AND RETRIEVAL PERFORMANCE: THE LOGICAL EVIDENCE

Dagobert Soergel

College of Library and Information Services, University of Maryland, College Park, MD 20742-4345, USA

This article presents a logical analysis of the characteristics of indexing and their effects on retrieval performance. It establishes the ability to ask the questions one needs to ask as the foundation of performance evaluation, and recall and discrimination as the basic quantitative performance measures for binary non-interactive retrieval systems. It then defines the characteristics of indexing that affect retrieval - namely, indexing devices, viewpoint based and importance-based indexing exhaustivity, indexing specificity, indexing correctness, and indexing consistency - and examines in detail their effects on retrieval. It concludes that retrieval performance depends chiefly on the match between indexing and the requirements of the individual query and on the adaptation of the query formulation to the characteristics of the retrieval system, and that the ensuing complexity must be considered in the design and testing of retrieval systems.

Indexing consumes the lion's share of database creation costs. Database designers need to know:

- whether indexing improves retrieval performance;
- whether the improvement is worth the cost; and
- what characteristics of indexing are important to achieve an improvement.

Many experiments have been conducted to answer these questions. But indexing characteristics and their effects on retrieval are so complex that they largely defy study in artificial test situations. Most experiments fail to account for important interactions among factors as they occur in the real world, and thus give results that mislead more than they enlighten, results that have little meaning for the assessment or improvement of operational retrieval systems. For example, an experiment studying the effect of exhaustivity of indexing must consider the factor of query formulation. Good searching varies the query formulation to adapt to the level of exhaustivity in indexing, in contrast, many experiments studying the effects of exhaustivity leave the query formulation unchanged in the name of experimental control and thus study the effects of exhaustivity under conditions of bad searching, a subject of little interest.

Providing guidance to system designers, indexers, and searchers requires taking a step back and exploring the many factors that bear upon the effects of indexing on retrieval performance, resulting in a framework for the study of these effects through both logical reasoning and empirical research. This is what this article sets out to do; it discusses the assessment of retrieval performance, defines the characteristics of indexing, and examines their effects on retrieval.

Information retrieval extends well beyond retrieval of bibliographic records and of text. Accordingly, we draw examples from other contexts as well, particularly from software reuse. (Software reuse requires the retrieval of existing software modules that can be incorporated into a new program to serve specified functions, such as sorting data or stemming English words.) We use **entity** or **item** as a generic term to include, for example, documents, program modules, and food products.

ASSESSING RETRIEVAL PERFORMANCE

*I can call spirits from the vasty deep.
Why, so can any I, or so can any man: but will they come
when you do call for them?
(William Shakespeare: Henry IV, Part 1, Act 3, Scene 1)*

At the foundation of all performance evaluation is a simple criterion: **Can one ask the questions one needs or wants to ask and get an answer with acceptable levels of successful retrieval and of distracting noise?** The remainder of this section gives variations on this fundamental theme.

RELEVANCE, PERTINENCE, UTILITY

The concepts of topical relevance, pertinence, and utility are central for good indexing and for assessing retrieval performance.

Topical relevance is a relationship between an entity and a topic, question, function, or task. A document is topically relevant for a question if it can, in principle, shed light on the question. "In principle" means that the document can do so for a person (or system) who knows the language of the document, has the background to understand the document, and is capable of processing the information transmitted by the document in relation to the question. "Shedding light on" means that the document provides information that either directly answers the question or is part of a premise set from which the question can be answered through a chain of inferences. The degree of relevance of a document depends on a number of factors: the amount of relevant information given; the strength of the relationship between the information given and the question - for example, the length of the inference chain from the information given to the answer; the strength of the contribution to the quality and/or surety of the answer; and perhaps other factors. (On a discussion of topical relevance, see Cooper [1971] and Wilson [1973].)

For a software module, the question of relevance is a question of applicability to a function to be carried out, such as sorting. A software module is topically relevant for a function if it can, in principle, assist in creating a program carrying out the function. "In principle" means that the software module can do so for a programmer who has the equipment to run the software module, the knowledge needed to fit the module in a larger whole, and the ability to modify the module if necessary. "Assisting" means that the module can be used as is or with modification; that it can serve as a model which facilitates creating a new module (for example, the module found could be recoded in another programming language); or that it can be incorporated into a larger program being written to carry out the function.

Pertinence is a relationship between an entity and a topic, question, function, or task with respect to a person (or system) with a given purpose. An entity is pertinent if it is topically relevant and if it is **appropriate** for the person, that is, if the person can understand the document and apply the information gained. (On pertinence see Kemp [1974].)

Utility. An entity has utility if it is pertinent and makes a useful contribution beyond what the user knew already. Utility might be measured in monetary terms ("How much is having found this document worth to the user?") (Cooper, 1973; Goffman & Newill, 1967). A pertinent document may lack utility for a variety of reasons; for example, the user may already know it or may already know its content.

PERFORMANCE MEASURES

Retrieval systems should be judged by what they are designed to do - retrieving relevant items and rejecting irrelevant items in response to a query. The correctness of retrieval decisions is measured by **recall**, the fraction of all relevant items correctly retrieved. Recall can be seen as the probability of an item being retrieved, given that it is relevant. The correctness of rejection decisions is measured by

discrimination, the fraction of all irrelevant items correctly rejected. Discrimination can be seen as the probability of an item being rejected, given that it is not relevant. The complement of discrimination is fallout, the fraction of all irrelevant documents incorrectly retrieved (fallout = 1 - discrimination). The definitions of recall and discrimination can be based on topical relevance, pertinence, or utility, possibly taking into account the degree of relevance (such as considering only highly relevant documents in the computation of recall or including moderately relevant documents as well).

The choice of the measures of retrieval performance to be used in logical analysis and empirical studies is important. Recall and discrimination are the two elemental, basic measures. The more commonly used measure **precision** (the fraction of relevant items in the items retrieved) is a composite measure, the joint result of several factors: The number of relevant items in the collection and recall together determine the number of relevant items retrieved; collection size and discrimination together determine the number of irrelevant items retrieved. Precision is a sensible measure of final answer quality from a user's point of view, but discrimination (or its complement, fallout) is the better measure for studying retrieval effects. Mooers (1959) uses recall and fallout. Goffman and Newill (1967) employ an analogy with the assessment of diagnostic tests and introduce recall (called sensitivity, the fraction of people having a disease testing positive) and discrimination (called specificity, the fraction of people having a disease testing negative). In a comprehensive review, Robertson (1969) convincingly argues the merits of fallout for measuring retrieval performance. The predominance of precision in spite of this can perhaps be explained by a bandwagon effect.

Much effort has been spent on defining a single measure of retrieval performance that could be used in comparing systems (Good, 1967; Swets, 1963). Such a single measure is meaningful only with reference to a specific request from a specific user in a specific situation. The answer to the question "which retrieval result is better" often depends on the user and her situation; different search requests have different requirements, and the quality of a search can be judged only in light of these requirements. Thus, a retrieval performance score for an individual query should be computed as a function of several criteria, each weighted according to search requirements as proposed by Bourne et al. (1961). A real-life system can be judged by the average of individual retrieval performance scores, with each search weighted by importance. This method favors flexible systems that let the searcher emphasize the search outcomes required in each search.

In addition to recall and discrimination there are other criteria to judge retrieval effectiveness, such as timeliness and novelty; for a discussion in the framework of bibliographic retrieval systems, see, for example, Soergel (1985). Beyond that, the support a system gives to a searcher in coming up with a good query formulation in the first place is critical for performance.

The discussion so far has been limited to binary retrieval systems: items are considered either relevant or not and are considered either retrieved or not. Now relevance is clearly a matter of degree, and systems are available that make retrieval a matter of degree by assigning each item a score. Such scores can be interpreted either as the degree of relevance predicted by the system or as the system-determined probability that the item is relevant on a binary scale; either way, the scores can be used to rank the items for examination by the user. The implications for measures of retrieval performance are discussed, for example, in King and Bryant (1971, ch. 2) and Swets (1963, 1969).

One further qualification is needed. The measures discussed assume the approach of a "one-shot" search: A query encompassing the entire information need is formulated, the search run, and the results evaluated. In real life, many searches are at least somewhat interactive: The searcher uses clues from entities retrieved to detect new search paths or to change the search topic, as exemplified most vividly in hypermedia systems. In this process, the searcher may accumulate the information needed in bits and pieces gleaned from documents encountered - the "berry-picking" approach to information retrieval (Bates, 1989). Interactive searches require additional criteria and measures.

While this article focuses on the effects of indexing on retrieval performance in a binary retrieval system as measured by recall and discrimination, the thinking presented also sheds light on evaluation of retrieval more generally.

INDEXING CHARACTERISTICS

Before we can talk about the effects of indexing on retrieval performance, we need concepts and vocabulary to talk about the characteristics of indexing that affect retrieval. This section reviews well-known indexing characteristics and defines some new ones. (In Sections Indexing Characteristics and Effects on Indexing Characteristics on Retrieval Performance see King & Bryant [1971, chs. 4 and 5].)

THE OVERALL APPROACH TO INDEXING

The approach taken to indexing can be expected to have a major impact on indexing quality and therefore on retrieval performance. Indexing can be request-oriented or entity-oriented (See "User-Centered Indexing" by Fidel [1994], Soergel [1985, ch. 14], or Cooper [1978]). In request-oriented indexing, the index language is built from a detailed study and logical analysis of user requirements and then serves as a communication device from user to indexer. The index language communicates to the indexer a conceptual framework to be used as a checklist in indexing. The idea is to maximize as far as possible the probability that a descriptor needed in searching is available in the index language and is used properly in indexing.

INDEXING DEVICES

There are structural and syntactical indexing devices. The most important structural device is the **hierarchy** of the index language. A well-structured hierarchy, preferably using facet arrangements where appropriate, provides a framework for the indexer and thus supports correct indexing; it is a prerequisite for request-oriented indexing. Hierarchy also supports searching. Associative relationships augment the hierarchy. Synonym relations constitute another structural device which is very important if the indexing vocabulary is not controlled.

The **degree of precombination** (precoordination) is another element of the index language structure (see Soergel, 1985, Chapters 14 and 15). A high degree of precombination makes indexing more difficult and thus may affect indexing correctness, if for no other reason than the increase in the number of descriptors in the index language; as we shall see, precombination also tends to degrade indexing specificity.

Two syntactical devices that are perhaps more important than generally assumed are links and role indicators/relators (see Körner [1985] for a review). **Links** express relationships between descriptors that are stronger than mere co-occurrence in the indexing of the same entity. They are particularly important in systems with a low degree of precombination. For example, a document about *The effects of alcohol dependence on experimentation with cocaine* would be indexed with *drug dependence: alcohol* and with *experimentation: cocaine*; the links prevent its retrieval in a search for *cocaine dependence*, which would use the query formulation *drug dependence: cocaine*.

Role indicators specify the role of a descriptor in the context of an entity representation. For example, Medline has the role indicators (called subheadings) *therapeutic use* and *adverse effects* to distinguish between the roles a chemical substance may play in a document, for example, *diabetes - drug treatment*, *insulin - therapeutic use*, *triglycerin - adverse effects*. Role indicators must be linked to the descriptor whose role they specify, lest confusion reign. **Relators** represent patterns of role indicator pairs (or triplets, etc), for example, *diabetes - treated by drug - insulin*.

Weights as an indexing device are discussed in connection with exhaustivity in the next section.

Exhaustivity of Indexing

An entity being indexed is relevant, in varying degrees, to a number of concepts. **Exhaustivity of indexing** is the extent to which these concepts are covered by the descriptors assigned to the entity. Exhaustivity has two components: viewpoints exhaustivity and importance exhaustivity.

TABLE 1 Importance exhaustivity (entity view)*.

High importance threshold Few descriptors, low exhaustivity	Low importance threshold Many descriptors, high exhaustivity
Use a descriptor only if the entity is definitely useful in a search for the descriptor.	Use a descriptor also if the entity might be useful in a search for the descriptor, even if the relationship is only tenuous, e.g. little space devoted to the descriptor topic.
Index only the main topics of a document.	Index also minor topics of a document.
Index only the main research method used.	Index all research methods used.
Index only the first ingredient of a food product.	Index all ingredients of a food product.
Index a program module with a function descriptor (such as <i>Sort or Reduce to stem form</i>) only if the module carries out the function.	Index a program module with a function also if the module could be modified to carry out the function, or if it suggests a good programming technique for writing a new module for the function.

* Sample indexing rules in a variety of systems.

A concept cannot be covered in indexing unless it is included in the index language. **Viewpoint exhaustivity** addresses the question: Are the facets or viewpoints useful for retrieval represented in the index language and thus available for retrieval? The degree to which this question can be answered with "yes" is viewpoint exhaustivity. For example, adding the three facets *level of difficulty*, *quality*, and *ideological orientation of the author* increases exhaustivity; so does adding the facet *research method used* in the indexing of research studies.

Importance exhaustivity address the question: What is the importance threshold for the assignment of descriptors as prescribed in the indexing rules? For the indexer considering an entity this question takes the form: Which of the concepts associated with this entity are important enough to warrant indexing? This question addresses the entity view of importance exhaustivity, as illustrated in Table 1.

The searcher considering a descriptor for a query formulation looks at the importance threshold from a different perspective: Will the descriptor find only documents that are centrally important or will it also find documents that just touch on the descriptor? This question addresses the descriptor view of importance exhaustivity. Exhaustivity - a system's indexing rules prescribing an importance threshold - may vary from subject area to subject area or even from descriptor to descriptor. For example, an information system for food products may prescribe indexing with *wheat* if it is the first or second ingredient but indexing with *mushroom* whenever mushroom is an ingredient, no matter how far down the list (since mushrooms pose a danger to food safety). What really matters in a search is not some general importance threshold but the importance threshold used for each of the descriptors that make up the query.

The rules given to the indexers define the level of exhaustivity intended in the system. Retrieval performance depends also on the application of these rules. Thus, to study the effects of exhaustivity (both viewpoint exhaustivity and importance exhaustivity) one must also consider the completeness of indexing, to be discussed below.

To test the effects of exhaustivity of indexing or to give guidance to a searcher in using a database, one must somehow measure exhaustivity. Unfortunately, measuring indexing exhaustivity is problematic.

Given two databases covering the same entity type (two databases covering documents, two databases covering software modules, etc.), we can determine which uses higher exhaustivity as described below (keeping in mind that database 1 may be more exhaustive in one subject area and database 2 in another). Thus, we can arrange several databases in a rank order based on their exhaustivity of indexing.

To compare the viewpoint exhaustivity of two databases, we can compare the index languages used with respect to the facets and individual descriptors included. To compare the importance exhaustivity, we can examine the indexing rules (such as the rules given in Table 1). We might even define an ordinal scale of exhaustivity values defined by such rules. But assigning a quantitative value to the importance threshold used in indexing is well-nigh impossible: any such value would be a fuzzy measure at best.

The average number of descriptors assigned to an entity in the database being studied is often used - somewhat naively - as a stand-in measure for exhaustivity. This would work if exhaustivity was the only determinant of the number of descriptors per document. But there are other factors: The properties of the entity being indexed, the degree of precombination, the correctness of indexing, and the indexing policy with respect to the assignment of descriptors that are broader or otherwise related to a "best-fit" descriptor. Table 2 shows these influences in detail. (See Maron [1979] for further discussion.)

From a descriptor view, we can compare the exhaustivity in two systems that have the same collection and use the same index language but differ in other parameters, such as indexer, amount of information used in indexing, amount of time used, and index language displays used. The system with more entities for a given descriptor is more exhaustive with respect to that descriptor.

Terminological note: The term indexing depth should be avoided: it is sometimes used to mean exhaustivity, sometimes specificity, and sometimes a combination of both.

Table 2. Determinants of the numbers of descriptors per entity.

	Influence on the number of descriptors per entity
Entity properties	At the same level of exhaustivity, a simple program module needs fewer descriptors than a complete software package. A long document often needs more descriptors than a short one. However, a 20-page journal article may need 15 subject descriptors at a medium level of exhaustivity, whereas an economics textbook needs just one: <i>Economics, general</i> (i.e. all subjects of economics covered)
Degree of precombination	The single precombined descriptor Methods of instruction for reading in elementary schools entails three elemental concepts: in a postcombination system the same topic requires three elemental descriptors, namely Methods of instruction, Reading, and Elementary schools. Thus, as a rule, the higher the degree of precombination, the fewer descriptors are required at a given level of exhaustivity. Therefore, comparing the exhaustivity of indexing of a given entity in two systems requires that all descriptors be reduced to elemental concepts.
Correctness of indexing Indexing rules	Incorrect descriptors increase the count without contributing to exhaustivity. Some systems require that the indexer add descriptors that are broader than or otherwise related to a "best-fit" descriptor; for example, a document on <i>clinical depression</i> would be indexed by that descriptor (which makes the best fit), but also by the broader descriptors <i>affective psychosis, psychosis, and behavioral and mental disorder</i> - three additional descriptors that do not add to exhaustivity.

Weights. Indexing weights differentiate descriptor assignments by importance: this is very useful in systems that use a low importance threshold (high exhaustivity) in indexing. Weights are commonly used in bibliographic databases that are distributed in both print and electronic formats, such as Medline and ERIC: the print index uses only highly weighted descriptors, but the electronic format uses all descriptors.

Specificity of Indexing

Specificity of indexing is the generic level at which the concepts assigned to the entity are expressed. Indexing by *clinical depression* is more specific than indexing by *psychosis*, indexing by *shell sort* is more specific than indexing by *sort*.

Comparing two systems with respect to specificity may be tricky. While one system may be consistently more specific than another, more often one system is more specific in one aspect, and the other system more specific in another aspect. For example, system A may index *diseases* specifically and *research methods* broadly, while system B indexes diseases broadly and *research methods* specifically.

As a practical matter, postcombination indexing is often more specific than precombination indexing. The three elemental descriptors *Methods of instruction: Reading: Second grade* provide more specificity in the grade level facet than the precombined descriptor *Methods of instruction for reading in elementary schools*. It is just not feasible to include all possible precombinations. Therefore, comparing the specificity of indexing in two systems requires that all descriptors be reduced to elemental concepts.

Intended specificity is the standard defined by the descriptors in a system's index language and by indexing policies. Actual specificity may fall short of the standard set by intended specificity due to indexing errors. An indexer may be unable to differentiate between specific descriptors, or information used in indexing may be insufficient for determining the best specific descriptor. Examples are given in the section on Effects of Indexing Specificity.

INDEXING CORRECTNESS

Indexing correctness or absence or indexing errs is of overriding importance for retrieval performance. Indexing is susceptible to two kinds of errors, two kinds of deviations from the indexing rules: errors of omission - a descriptor that should be assigned is omitted: and errors of commission - a descriptor that should not be assigned is nevertheless assigned. Omitting a correct descriptor and assigning a broader or narrower or related descriptor instead is a special kind of error that is at once an error of omission and commission. In such a case, the searcher can compensate for the error of omission by including related descriptors in the query formulation, and the error of commission is mitigated since an item retrieved by the incorrect, but related, descriptor has at least some relevance. In the measures defined below, some credit should be given for descriptors that are half-correct, but also some blame, since the same descriptors are also half-wrong. To further complicate matters, highly weighted descriptors should be given more weight in computing indexing correctness.

To determine indexing errors, one must know for each entity which descriptors should and should not be assigned. In reality, things are not that clear-cut. There may well be differences between good indexers, but we assume that correctness of indexing can be determined, for example, through a consensus of several good indexers and knowledgeable users.

With this caution in mind, we define the following indexing correctness. **Completeness of indexing** relates to the presence of the correct descriptor assignments, the absence of errors of omission. We can look at the completeness of indexing from two points of view. Considering an **entity** and the descriptors assigned to it, we can ask: 'Of the descriptors required for this entity by the rules of the system, how many are actually assigned?' This question addresses completeness for each entity taken individually - the entity view of completeness. As a formula:

Completeness of indexing (entity view)

no. of descriptors that are assigned
correctly to the entity

no. of all descriptors that should be
assigned to the entity

A searcher considering the use of a **descriptor** for a query formulation is more interested in a different question: Of all the entities that should be indexed by the descriptor, how many are actually so indexed? This question addresses completeness from the point of view of a descriptor as it appears in the whole collection - the descriptor view of completeness (King & Bryant [1971,p.139], "Indexing Accuracy"). As a formula:

Completeness of indexing (descriptor view)

no. of entities to which the
descriptor is correctly assigned
no. of all entities to which the
descriptor should be assigned

The descriptor-view measure is more difficult to determine, but it is more useful for predicting retrieval performance; it is directly related to recall.

The **purity of indexing** relates to the absence of erroneous descriptor assignments, the absence of errors of commission. Purity of indexing can also be viewed from an entity taken individually or from a descriptor as it occurs (or does not occur) in the collection. The formulas are:

Purity of indexing (entity view)

no. of descriptors that were
correctly rejected for the entity
no. of all descriptors that should
have been rejected for the entity

Purity of indexing (descriptor view)

no. of entities for which the descriptor
was correctly rejected
no. of all entities for which the
descriptor should have been rejected

Again, the descriptor-view measure is more difficult to determine but it is more useful for predicting retrieval performance; it is directly related to discrimination.

There are other measures that address indexing errors of commission. We chose purity of indexing because it is a positive measure; the higher its value, the better the indexing. Its complement, impurity of indexing, would count incorrect descriptor assignments rather than descriptor assignments correctly rejected; in some ways, that might be more intuitive, but the parallel to completeness would be lost. Impurity of indexing (descriptor view) is directly related to fallout. Other intuitive measures would be the fraction of all descriptors assigned to an entity that are correct or incorrect, respectively, but the relationship of these measures to retrieval performance is less clear.

Indexing correctness must be measured in terms of the rules of the system. A document may be indexed with three elemental descriptors in system A and be 100 percent complete (system A uses very low exhaustivity and requires only three descriptors) while being indexed with 20 elemental descriptors in system B and still being only 50 percent complete (system B uses high exhaustivity and requires 40 descriptors). A descriptor correctly assigned in a high-exhaustivity system may be erroneous in a low-exhaustivity system.

INDEXING CONSISTENCY

Indexing consistency is not important in itself; it has been used as an indicator of indexing correctness, but that is problematic. Indexing can be consistently wrong; all indexers might miss an important implication of a document and thus omit an important descriptor. On the other hand, high indexing correctness results in high consistency (two indexers achieving indexing completeness and purity of 1 are also entirely consistent); thus, high consistency is a necessary, but not sufficient, condition for high correctness. The situation is parallel to measurement: high reliability is a necessary, but not sufficient, condition for high validity.

Indexing correctness and indexing consistency are both measures of agreement. Indexing correctness is measured as the asymmetric agreement of the descriptors assigned by the indexer with the descriptors that should be assigned. Indexing consistency can be measured by any of the many measures for symmetric or asymmetric agreement between two indexers (interindexer consistency) or between two indexing sessions by the same indexer (intraindexer consistency). The considerations on descriptor relationships and descriptor weights from the previous section - Indexing Correctness - apply here as well. The following is an example of a symmetric measure:

Indexing consistency (entity view)
no. of descriptors assigned to the
entity by both A and B
no. of descriptors assigned
to the entity by A or B

Indexing consistency (descriptor view)
no. of entities to which the descriptor
was assigned by both A and B
no. of entities to which the descriptor
was assigned by A or B

(Descriptor-view measure defined in King & Bryant [1971, p. 138].)

EFFECTS OF INDEXING CHARACTERISTICS ON RETRIEVAL PERFORMANCE

There are a few simple truths that make the relationship between indexing characteristics and retrieval performance extraordinarily complex and very hard to sit through experiments.

- Important determinants of retrieval performance are not universal but idiosyncratic to the query at hand: Does the index language include the descriptors needed to express the query topic? Do the indexers' judgments in applying these descriptors match the requirements of the query? Does the index language include the hierarchical relationships useful for processing the query? In short, indexing performance depends on the match between indexing characteristics and the requirements of the individual query.
- The quality of searching plays a major role both in exploiting the strengths of indexing to the fullest and in compensating for its weaknesses. It is good search practice to adapt the query formulation to the indexing environment, for example, to exhaustivity of indexing; this is the only way to ensure the best retrieval possible under the constraints of the given indexing character. This principle creates a quandary for experiments that try to compare several retrieval systems; if such an experiment uses the same query formulation with all of them (to hold this factor constant), it measures retrieval performance under conditions of bad searching.
- Indexing characteristics are not the only determinant of retrieval performance; the retrieval mechanism also plays an important role. The searcher can do a lot more with a powerful computer search system than with a printed index or card catalog.

These truths are illustrated again and again in the analysis that follows.

EFFECTS OF THE APPROACH TO INDEXING

The approach to indexing has a large effect on the availability of just the right descriptors needed for searching and on the correctness of indexing; thus it can be expected to have a major impact on retrieval performance. Request-oriented indexing is designed to increase dramatically the ability to ask the questions one needs or wants to ask. Unfortunately, there is only one rather small study investigating this important variable, with encouraging but preliminary results (Pejtersen, 1980, 1986; Pejtersen & Austin 1983, 1984).

EFFECTS OF INDEXING DEVICES

Effects of Hierarchy. Hierarchy provides a framework for the indexer and thus has positive effects on indexing correctness, which in turn improves retrieval. Hierarchy also has a direct effect on searching. It provides a framework for the searcher in formulating the query and thus supports choosing the most appropriate descriptors. The structure may even assist the user in thinking about the problem and discovering ramifications and new aspects.

In the search itself, hierarchy is the basis for **inclusive searching** (Medline EXPLODE, Predicasts CASCADE). In inclusive searching, a query descriptor retrieves entities indexed by any narrower descriptor as well; for example, an inclusive search for *psychosis* finds as well documents indexed by any specific *psychosis*, such as *schizophrenia* (including *paranoid schizophrenia*, *schizoaffective disorder*, etc.) and *affective psychosis* (including *manic disorder*, *depression*, etc.). Inclusive searching applies the knowledge encapsulated in the index language hierarchy to provide a very powerful search tool that boosts recall. Inclusive searching is particularly important in systems that use high specificity of indexing or a high degree of precombination.

Effects of Precombination. Precombination comes in very handy when there is a precombined descriptor that matches the query topic. In a printed index, where searching for a combination of descriptors is impractical, a precombined descriptor matching the query topic makes searching feasible. In a computer search system, a precombined descriptor help avoid spurious combinations; for example, combining the elemental descriptors *drug dependence* AND *cocaine* without links would retrieve, through spurious combination, a document on *The effect of alcohol dependence on experimentation with cocaine*; using the precombined descriptor *cocaine dependence* would prevent the erroneous retrieval. Spurious combinations can also be prevented through the use of links.

Precombination can also create difficulties in retrieval. If *alcohol dependence*, *cocaine dependence*, *marijuana dependence*, and *nicotine dependence* are all precombined descriptors, they must all be included in a search for *drug dependence*. This does not present a problem if the search system supports inclusive searching and if the hierarchy includes the relationships *drug dependence* NT *alcohol dependence*, etc.; unless both conditions are fulfilled, the searcher must compensate for the deficiency by including the precombined descriptors in the query formulation, lest recall suffer. For example, even though Medline supports inclusive searching, a Medline search for *eye, inclusive* does not retrieve all documents on *eye neoplasms* since *eye neoplasms* is not a narrower term of *eye* and the database contains many documents indexed by *eye neoplasms* but not by *eye*.

Effects of Links and Role Indicators. Links and role indicators can be used to formulate a more discriminating query, but their use makes for another opportunity for mismatch between indexers and searchers, and thus recall may drop. Many links are very obvious and easy to use consistently, such as the link between *drug dependence* and the specific drug. Role indicators are more complex. Rules can be overly restrictive, such as allowing only one role indicator for a descriptor, while in reality a concept often plays several roles in the context of a document; for example, in Medline indexing, a prescription drug is often shown in the two roles *therapeutic use* and *adverse effects*. The semantic relationships between role indicators must be considered in searching to counteract a mismatch of interpretation between indexer and searcher; for example, the roles *acted upon* and *result* have a certain semantic affinity that may lead to the indexer using one while the searcher would search under the other. Thus, it is prudent to search for *acted upon* OR *result*. Early experiments gave role indicators a bad name (e.g., Sinnett, 1963). However, the

performance degradations reported reflect more on poorly designed systems operated by poorly trained people than on the merits of the concept. Montague (1965) gives a balanced account. Role indicators may be particularly useful for queries that combine very broad concepts in specified roles.

EFFECTS OF INDEXING EXHAUSTIVITY

common wisdom has it that indexing exhaustivity increases recall at the cost of discrimination and that specificity increases discrimination at the cost of recall. Common wisdom has it wrong half of the time.

Higher **viewpoint exhaustivity** extends the kinds of questions that can be asked of a system. It also affords the searcher an opportunity to achieve higher discrimination, possibly but not necessarily at the cost of decreased recall. Assume a system adds *research method used* to the viewpoints indexed. In a search for *longitudinal of attitudes toward drunk driving using the panel method* the searcher can now add a research methods descriptor, *panel study*, to increase discrimination. The amended query formulation may also fail to retrieve some relevant documents, depending on the importance exhaustivity and the indexing completeness associated with the descriptor *panel study*. (Indexing completeness might be low if the indexers are not sufficiently trained in research methods to recognize the use of the panel method.) When viewpoints such as *reading level* or *program complexity* are indexed, discrimination with respect to pertinence can be increased. Adding viewpoints to a query formulation is strictly an option; if the searcher expects that this option would reduce recall and recall is at a premium, she may elect not to use that option.

The effects of **importance exhaustivity** are more complex. When the query formulation is held constant, higher exhaustivity delivers the same or higher recall with the same or lower discrimination, depending on the situation. Consider a search for *depression* by two different users: (1) a physician who wants to brush up on the latest developments in *depression*; and (2) a medical researcher who wants to do a thorough review of all aspects of *depression*. Both users run the same search and get the same results, but their differing requirements lead to a different interpretation of these results as illustrated in Table 3.

High importance exhaustivity can sometimes be exploited to achieve higher discrimination, since it affords the searcher an opportunity to reformulate the query by adding an additional restrictive criterion. Consider a search on *alcohol dependence and depression*; a document is judged relevant if it concentrates on *alcohol dependence* and at least mentions *depression*. With low exhaustivity the query should be just *alcohol dependence*; qualifying with *depression* would reject so many relevant documents as to be unacceptable. The documents found must be screened for mention of *depression*. With high exhaustivity and weights, the query could be narrowed to *alcohol dependence* (highly weighted) AND *depression* (any weight). This would boost discrimination and detract little from recall; the only documents missed would be borderline documents where the indexer thought the mention of *depression* too insignificant while the user judged the document relevant. In each case, the query formulation was adapted to achieve the best retrieval possible, given the level of importance exhaustivity. In bibliographic and other text-based systems, free-text searching may be used to compensate for low indexing exhaustivity, provided the search concept is represented explicitly by a word or phrase in most relevant documents.

TABLE 3. Effects of importance exhaustivity for different user requirements

User	Low exhaustivity	High exhaustivity
A physician who wants to brush up on the latest developments in <i>depression</i> .	Retrieves most of the documents the physician needs without too much noise. The noise consists of borderline documents that the indexer judged quite important for depression but that the user finds irrelevant. The missed documents are borderline documents that the indexer did not judge important enough but that the user would find relevant. Good recall, good discrimination. Good importance threshold match: The indexer uses high, the user requires high.	Retrieves the same documents, plus many more that just touch on depression and thus are not of interest to the physician. A few of the added documents may be relevant - borderline documents that are judged differently by the indexer and the user. Quite good recall, poor discrimination. Importance threshold mismatch: The indexer uses low, the user requires high.
A medical researcher who wants to do a thorough review of all aspects of <i>depression</i> .	Misses many of the documents the researcher needs and rejects almost all irrelevant documents. Poor recall, quite good discrimination. Importance threshold mismatch: The indexer uses high, the user requires low.	Retrieves many more documents, most of them relevant for the researcher. Good recall, good discrimination. Good importance threshold match: The indexer uses high, the user requires high.

Indexing weights allow the searcher to choose for each query descriptor the level of importance exhaustivity best suited to the user's requirements, as illustrated in the preceding example. In a high-exhaustivity system without weights, the query formulation would be *alcohol dependence AND depression*, retrieving documents in which both disorders are mentioned, including documents where alcohol dependence is just mentioned, not dealt with intensively as required by the user. Thus, this query would find both more and less, and the effect on recall and discrimination depends on what is in the collection.

EFFECTS OF INDEXING SPECIFICITY

The effects of indexing specificity depend on the specificity of the search. A specific search can take advantage of specific descriptors to increase discrimination, but for a broad search specific indexing does not help and may even be harmful.

The effects of specific indexing on a specific search depend on the correctness of indexing: specific descriptors make the search more vulnerable to indexing errors. If the indexing is correct, using a specific descriptor hurts recall very little, if at all. But if the indexers are often unable to determine the proper specific descriptor and resort to assigning a broader descriptor instead, specific descriptors become less reliable. The following examples, from retrieval systems dealing with different kinds of entities, illustrate the point. A document relevant for *x-ray film tomography* may have been indexed, erroneously, under the broader descriptor *film radiography* or by the neighbor descriptor *scanography*. A job seeker qualified for an open job as *furnace operator* may have been indexed by the broader descriptor *foundry worker*. A data set relevant to the study of attitudes toward *power* may have been indexed by the neighbor term *authority*. A patient suffering from *depression* may have been diagnosed under the broader descriptor *psychosis*. In all these cases, the broader descriptor retrieves the relevant entity, but the specific descriptor misses it due to the indexing error. With inclusive searching, the searcher can always search under the broader descriptor inclusively to increase recall at the cost of discrimination.

The effect of specific indexing on a broad search depends on the capabilities of the search system. Assume a broad search on *psychosis*; relevant documents may be indexed by *psychosis* or any of its narrower descriptors, such as *affective psychosis*, *depression*, etc. As long as the index language has a well-developed hierarchy and the search system supports inclusive searching, this is not a problem and neither search effort nor retrieval performance are affected. But when inclusive searching is not provided, the searcher looking for a broad descriptor must remember to include in the query formulation all the narrower descriptors seen from the hierarchy - quite laborious and a drawback of specificity in the environment of poor retrieval software or of card catalogs and printed indexes. When the index language does not have a good hierarchy, things are even worse: The searcher must think of all the narrower descriptors, possibly consulting reference tools. Since the searcher cannot be expected to come up with a complete list of narrower descriptors, recall suffers.

EFFECTS OF INDEXING CORRECTNESS

Indexing correctness is a major determinant of retrieval performance. Assume a one-concept query that can be expressed well by a descriptor in the index language: assume furthermore that the importance threshold used by the indexers matches the user's requirement. In this situation recall is equal to the completeness of indexing (descriptor view), and discrimination to the purity of indexing (descriptor view). With correct indexing, the searcher can rely on the descriptor assignments and feel confident about retrieval results. This is particularly important in the search for a concept that would be hard to search using a free-text approach.

The following examples, taken from the *Alcohol and Other Drugs Thesaurus* (1993), illustrate this point. The thesaurus includes the descriptor *acculturation* because of the interest in this topic. A document on the *introduction of alcohol into the community of the Hare Indians and the integration of drinking behavior into their culture* must be indexed by *acculturation*, or it will be lost to retrieval under this topic or the broader topic *socio-cultural change* - an indexing error of omission leads to low recall. (The abstract of the document does not mention *acculturation* or *change*, thus a free-text search would not help. On the other hand, a free-text search for *Hare Indians* might be quite successful.) The thesaurus also includes the descriptor *dual diagnosis* for the condition of simultaneous *drug dependence* and another *behavioral or mental disorder* because this is a "hot topic." Indexers must be vigilant to spot the co-occurrence of any kind of *drug dependence* with any other *behavioral or mental disorder in the same person* and then use *dual diagnosis* (co-occurrence of appropriate terms in the same abstract is **not** enough). As a last example, consider the document on *The effects of alcohol dependence on experimentation with cocaine*; it must be found under the descriptor *gateway drugs*. Again, if the indexers' vigilance fails, retrieval losses will occur. Conversely, "over-indexing" with a descriptor due to lack of proper understanding hurts discrimination. These examples illustrate that correct indexing requires indexers who are well trained and familiar with the subject matter and the needs of the users as reflected in the index language.

In practice it is difficult and laborious to measure the correctness of indexing; experiments to study its effects are, therefore, at the borderline of feasibility. Their focus would need to be not so much on the positive effects of indexing correctness, which is self-evident, as on the question to what extent lack in indexing correctness can be overcome in searching.

EFFECTS OF INDEXING CONSISTENCY

Indexing consistency affects retrieval performance only indirectly, through its possible effect on indexing correctness, which was discussed in the section on Indexing Consistency.

INDEXING CHARACTERISTICS, RECALL, AND DISCRIMINATION

Indexing characteristics are often divided into those that support recall and those that support discrimination. However, as the foregoing analysis shows, many indexing characteristics support both, depending on the situation.

A recall device can often be exploited for a query reformulation that boosts precision. The following example illustrates this for the recall device **inclusive searching**. Assume a search on *depression in elderly cancer patients*. There are three search concepts, *depression*, *elderly*, and *cancer*. *Cancer* has many, many narrower descriptors; without inclusive searching, they would all have to be ORed to express the third search concept - not a feasible option. Free-text searching does not help, since the text of many relevant documents contains a term for a specific kind of cancer but not the word *cancer* itself. So the searcher goes with *depression AND elderly* for a low discrimination answer and then screens the documents found for mention of any form of cancer. With inclusive searching, the third search component can be expressed as *cancer, inclusive*, resulting in a search with high discrimination.

Conversely, a discrimination device, such as role indicators, often makes it feasible to ask a question that otherwise would return so many entities as to be unrealistic. In that case one can say that recall was increased from 0 (the value if the question is not even asked) to whatever can be achieved with a reasonably discriminating query.

CONCLUSION

The discussion of the characteristics of indexing and of the mechanisms through which they affect retrieval performance has demonstrated the complexity of the interactions not only among the indexing characteristics themselves but also with the characteristics of the search system used. To make matters worse, many of the variables defy quantitative measurement.

This situation is not made for classical experiments where all variables except the variable under study are held constant or manipulated to study their effects in a carefully controlled way. It may be unwise to accept the results of an experimental study in this area before the study design has been scrutinized to ensure that all the effects and interconnections discussed here have been considered and that all the pitfalls have been avoided. This area may be more suitable for qualitative research methods - methods that take a holistic approach and are sensitive to the special circumstances and total context of individual cases. Qualitative studies may not deliver the clear-cut and "safe" quantitative results some look for, but they may provide insights into the functioning of retrieval systems - perhaps tentative, perhaps uncertain, perhaps qualified - but insights close enough to reality to be useful for design.

Some design conclusions can be drawn from the logical analysis provided here. Best overall retrieval results can be achieved with systems that are flexible, systems that let the searcher emphasize the search outcomes in accordance with the requirements of each search. Such flexibility can be achieved through the following combination of features:

- An index language that covers all viewpoints needed in retrieval and arranges its descriptors in a well-structured hierarchy.
- Links and role indicators.
- Exhaustive indexing (low importance threshold) with weights.
- Specific indexing and inclusive searching.
- Well-trained indexers who are sufficiently familiar with the subject matter and with user needs as reflected in the index language to make the judgments necessary for correct indexing.

The discussion of retrieval effects revolved around an important principle: Retrieval performance is a function of the agreement of the judgments of two parties, the indexer and the user, with respect to the viewpoints used in indexing, the general importance threshold used, and the specific judgments regarding the importance of a specific descriptor for a specific entity. The goal of indexing must be to maximize that agreement.

This principle can be stated more generally. Information retrieval is about meaning. While we can in many cases get at meaning through statistical and syntactic/semantic processing, in many other cases - perhaps the more important ones - we cannot, and human judgment - no matter how often it is maligned as subjective - must step in.

ACKNOWLEDGMENTS

I thank Raya Fidel for her helpful comments and Lois Lunin for her guidance in making this a much more readable article.

REFERENCES

- The alcohol and other drugs thesaurus. A guide to concepts and terminology in substance abuse and addiction.* (1993). Rockville, MD: National Institute on Alcohol Abuse and Alcoholism and Center for Substance Abuse Prevention.
- Bates, M. J (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13, 407-424.
- Bourne, C. P., Peterson, G. D., Lefkowitz, B., & Ford, D. (1961). Requirements, criteria, and measures of performance for information storage and retrieval systems. Stanford Research Institute Project Report No. 3741 (AD 270942).
- Cooper, W S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7, 19-37.
- Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. Part I. Philosophy. *Journal of the American Society for Information Science*, 24, 87-100.
- Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. Part II. Implementation of the philosophy. *Journal of the American Society for Information Science*, 24, 413-424.
- Cooper, W S (1978). Indexing documents by Gedanken experimentation. *Journal of the American Society for Information Science*, 19, 107-119
- Fidel, K. (1991). User centered indexing. *Journal of the American Society for Information Science*, 42, 572-576
- Goffman, W., & Newill, V. A (1967). A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval*, 3, 19-25
- Good, I.J. (1967). The decision-theory approach to the evaluation of information-retrieval systems. *Information Storage and Retrieval*, 3, 31-34.
- Kemp, B.D. (1974). Relevance, pertinence and information system development. *Information Storage and Retrieval*, 10, 37-47.
- King, D.W. & Bryant, E.C. (1971). *The evaluation of information services and products*. Washington, DC: Information Resources Press.
- Körner, H.G. (1985). Syntax und Gewichtung in Informationssprachen: Ein Fortschrittsbericht über präzisere Indexierung und Computer-Suche. [Syntax and weighting in information languages: A state-of-the-art report on more precise indexing and computer retrieval.] *Nachrichten für Dokumentation*, 36, 82-100.
- Maron, M.E. (1979). Depth of indexing. *Journal of the American Society for Information Science*, 30, 224-228.
- Montague, B. A. (1965). Testing, comparison, and evaluation of recall, relevance, and cost of coordinate indexing with links and roles. *American Documentation*, 16, 201-208.

Mooers, C.N. (1959). *The intensive sample test for the objective evaluation of the performance of information retrieval systems*. Cambridge, MA: Zator Co.

Pejtersen, A.M. (1980). Design of a classification scheme for fiction based on an analysis of factual user-librarian communication, and use of the scheme for control of librarians' search strategies. In: O. Harbo & L. Kajber (Eds.), *Theory and application of information research* (pp. 146-159). London: Mansell.

Pejtersen, A.M. (1986). Design and text of a database for fiction based on an analysis of children's search behavior. In: P. Ingwersen, L. Kajber, & A.M. Pejtersen (Eds.), *Information technology and information use: Toward a unified view of information technology* (pp. 125-145). London: Taylor Graham.

Pejtersen, A.M., & Austin, J. (1983). Fiction retrieval: Experimental design and evaluation of a search system based on users' value criteria. Part 1. *Journal of Documentation*, 39, 230-246.

Pejtersen, A.M., & Austin, J. (1984). Fiction retrieval: Experimental design and evaluation of a search system based on users' value criteria. Part 2. *Journal of Documentation*, 40, 25-35.

Robertson, S. E. (1969). The parametric description of retrieval tests. *Journal of Documentation*, 25, 1-27.

Sinnett, J.D. (1963). *An evaluation of links and roles used in information retrieval*. MS thesis. Wright-Patterson Air Force Base, OH: Air Force Institute of Technology: Dec. 1963. 299p. (AD 432189).

Soergel, D. (1985). *Organizing information. Principles of data base and retrieval systems*. Orlando, FL: Academic Press.

Swets, J.A. (1963). Information retrieval systems. *Science*, 141, 245-250.

Swets, J.A. (1969). Effectiveness of information retrieval methods. *Journal of the American Information Society for Information Science*, 20, 72-89.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, 9, 457-471. ©

SOERGEL: This paper was first published in *Journal of the American Society for Information Science*. JASIS 45(8) 589-599, 1994 (ISSN 0002-8231) and has been reproduced here with the kind permission of John Wiley & Son, Inc.

REFLECTIONS ON TREC

Karen Sparck Jones

Computer Laboratory, University of Cambridge, New Museums Site,
Pembroke Street, Cambridge CB2 3QG, U.K.
E-mail: sparckjones@cl.cam.ac.uk

Abstract

This paper discusses the Text REtrieval Conferences (TREC) programme as a major enterprise in information retrieval research. It reviews its structure as an evaluation exercise, characterises the methods of indexing and retrieval being tested within it in terms of the approaches to system performance factors these represent; analyses the test results for solid, overall conclusions that can be drawn from them; and, in the light of the particular features of the test data, assesses TREC both for generally applicable findings that emerge from it and for directions it offers for future research.

1. INTRODUCTION

TREC is a new ballgame for IR research and development. It is richly interesting in itself, can be expected to have many consequences for IR study and practice, and should be noticed and built on by everyone interested in automatic indexing and retrieval. We can learn both from what it has done and what it has not done, and thus focus future research.

My purpose here is to review TREC as a whole, considering it both in its own right and in the context of IR R&D over time. I shall, further, examine TREC both at the object level of its results for indexing and retrieval, and at the meta level of test methodology in IR. This paper, as a review, is intended to complement Harman's factual accounts in Harman (1994a,b), and has to be supplemented by the individual papers, the questionnaire responses and, most obviously, by the tabulated test results given in the TREC Proceedings (Harman, 1993, 1994a). Thus, my aim is to draw some general conclusions about the significant findings in TREC, and about the directions to pursue in the future: What have we found out? Where should we go next?

But I must emphasise that this review is a personal one constituting a first attempt at an assessment of a large and continuing enterprise. In particular, in trying to draw conclusions about what the work to date has shown, I have been forced to simplify, so my judgments may be crude and controversial. My presentation of results and findings should thus be seen as a snapshot, successful if leading not only to further, closer analysis of the existing materials, but also to further, more discriminating tests.

1.1 Paper perspectives and organisation

It is possible to assess TREC from several different points of view. I shall start by noting my frame of reference for the TREC programme as an officially organised activity, and indicate what I see as the major relevant features of the TREC enterprise as whole. The TREC programme has specific characteristics from the IR point of view, constituting what I call the TREC case, both providing opportunities for and imposing constraints on retrieval itself. Thus, TREC has its particular motivation,

leading to choices of test data and evaluation strategy with material effects on the inferences to be drawn from the TREC findings. Although some of the object-level TREC data characteristics are broad ones that would also hold in other applications, some are more particular, so their external implications may be less clear. The actual retrieval results obtained in the TREC tests have thus to be assessed correspondingly carefully for their wider application. Again, the TREC programme involves, at the meta level, a strongly shaped evaluation methodology relying on an evaluation paradigm that has to be judged both for its own soundness and for its pertinence to any wider application of the TREC findings. At the same time, TREC has a generic context, in the current state of operational IR practice and IR research, that has to be borne in mind in reviewing TREC. Finally, the TREC effort itself, represented by the work and tests done by the participants, involving many different teams doing very different things, has not only more obvious implications for the range or solidity of the TREC findings, but also other interesting features. Thus, it is possible not only to ask what is old about TREC, in the sense of what past findings it endorses and helps significantly to establish, but also what is new about TREC, in its concerns as well as its findings.

In my analysis of the TREC effort and discussion of findings I shall make use both of the well established notion of *performance factors* affecting IR, and of the *environment/system* contrast and the test methodology to which it leads, involving a decomposition generating a *grid* design with cells representing *runs* each for some request set and document set (cf. Galliers & Sparck Jones, 1993). The environment/system contrast distinguishes factors according to whether they are environment *variables* with *values* that are fixed for any given set of runs for which system parameter settings are altered or, alternatively, for which system *parameter settings* are held constant through environment changes. As this suggests, the implications of this methodology are that reliable and informative tests require modifications to factors in both grid dimensions that are also subject, in a third dimension reflecting a kind of evaluation meta environment, to the use of different performance *gauges*.

My review is organised as follows. The next section, 2, describes the TREC case, supplying the background for the subsequent description and assessment of the research done. My descriptive analysis of the TREC work first lays out, in Section 3, the performance factors that have been investigated, and then, in Section 4, summarises the actual test *results* from this point of view, using a series of factor questions to structure the analysis. In Section 5 I present what I see as the main *findings* that emerge from the work when the experimental results are assessed in the context of the TREC data and evaluation framework, and in Section 6 draw my conclusions about the future IR work that needs to be done.

1.2 Terms of reference

I shall use the term 'TREC' as a global label for the succession of Text REtrieval Conferences, each covering a preparation and test *phase* terminating in a reporting workshop. TREC-1 was in 1992, TREC-2 in 1993, TREC-3 is underway with its meeting in November 1994, and TREC 4 is already being planned. TREC meetings are essentially restricted to participants, but there are published Proceedings (Harman 1993, 1994a). As already indicated, it is convenient under the TREC umbrella term to distinguish the official organisation of the TREC programme from the actual work effort put in by participants, and also to recognise that something that can be called the TREC enterprise is developing that subsumes not only programme and work, but a range of supporting or derivative activities. SMART at Cornell and elsewhere is an earlier analogue but, as will be considered more fully later, the scale of the TREC tests, size and heterogeneity of the TREC community, and the fact that the teams are working concurrently in a short time scale, mean that TREC has special properties and, hence, effects on the field.

As TREC-1 was so novel, while the time available for TREC-2 was too short for extended experiment, it is sensible to treat TREC-1 as a debugging phase with consolidation in TREC-2. I shall therefore take TREC-2 as representative of the work done and progress made so far and thus, unless explicitly indicating otherwise, be referring in using the term "TREC" to the tests done in TREC-2 and reported in the TREC-2 Proceedings.

2. THE TREC CASE

Although TREC falls within a well established experimental IR tradition, there is novelty in the stringency with which these experimental conventions are being applied, in the nature and scale of the test material, and as just mentioned, in the size of the participating community. So though the TREC operational details are fully described in Harman (1994b), I shall briefly summarise the main features of the enterprise for convenient reference here, and comment on points that are particularly important for my attempt to assess the TREC contribution to IR. This is also desirable because, although the TREC meetings are restricted, all the materials, both data and results, will be made publicly available and may – indeed ought to – be exploited in further research building on the TREC work, just as test collections have been in the past. These new materials are, however, much richer than the old ones, and also have the great advantage, first, of being in full machine-readable form and, second, of including output search results as well as the input test data.

I shall characterise the TREC case with respect to its *motivation*, *test data*, *evaluation paradigm*, and *generic context*.

2.1 TREC motivation

It is important in assessing TREC to take account of its sponsoring motivation and situation. The stimulus for TREC has come from U.S. government organisations concerned with intelligence operations and R&D sponsorship, notably (D)ARPA, responding, on the one hand, to the vast growth in communications (parties, events, and matter) and, on the other, to the development of computing, language, and information processing technologies that make it possible to attack real-world rather than laboratory problems. ARPA has been promoting competitive task programmes since the mid 1980s, first on speech recognition (SR), subsequently in spoken language systems (SLS) and message understanding (MUC), and more recently in machine translation (MT) and combined message retrieval and information extraction (TIPSTER-subsuming TREC and MUC type tasks). (For a convenient overview see the materials in Bates, 1993.)

This background is important because it has established a style of operation coloured by the properties of the initial SR case and affecting the implicit goals and actual data for the TREC case. Thus, whereas the IR community has its conventional evaluation methodologies (unlike the NLP community, which has had to make a major effort to develop evaluation strategies), the background to TREC leads to a pressure to perform, and specifically to win, in a fairly crude way that is not necessarily beneficial either to the TREC sponsors' own long-term goals or to the field as a whole. The SR case is especially dangerous because performance there can be simply but uncontroversially defined using single-number measures. The established IR mainstream evaluation style, in fact illustrated by TREC, naturally fits the TREC programme *zeitgeist*, but nevertheless has its limitations, as considered later.

2.2 TREC data

The salient, relevant points about this are as follows (for more detail see Harman, 1994b). The document set is large, over 1 million items, though it is made up of three large blocks used separately and not treated as a single whole. The material is from seven sources, individually varied and collectively various in topics and genres, though with much news story material. In general, there are documents from each source in each block, though the material from any source is not necessarily a full or correctly ordered set. The documents are predominantly full text, but some sources do not have rich headers of the kind familiar from journal articles, or even any significant headers at all.

The request set totals 150, but has only been used as subsets. The TREC requests, or *topics* in TREC parlance, have very distinctive properties that need to be taken into account in any attempt to draw conclusions for wider application from the TREC tests. From this point of view it is also useful to bear in mind the distinction between a user's information *need*, a literature *request*, and a search *query*. Information needs are often not fully or explicitly expressed, especially in writing; and historically there

has been a shift from the old-fashioned case where an initial informal natural language request is replaced by a formal boolean query, perhaps using controlled language terms, to the case where request and query are much closer because the latter is a simple list of terms, derived from the former, that is applied in searching to deliver a ranked output. In accordance with common IR convention, and also to avoid confusion with other uses of the word "topic," I shall normally refer to the TREC topics as requests. However, as the example shown in Fig. 1 makes clear, they are long and detailed, and much more elaborate than requests usually are, with elements of both need statement and query. The TREC requests have been very carefully for-

Number: 051

Domain: International Economics

Title: Airbus Subsidies

Description:

Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.

Narrative

A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

Concept(s):

- 1 Airbus Industrie
- 2 European aircraft consortium, Messerschmitt-Boelkov-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
- 3 federal subsidies, government assistance, aid, loan, financing
- 4 trade dispute, trade controversy, trade tension
- 5 General Agreement on Tariffs and Trade (GATT) aircraft code
- 6 Trade policy Review Group (TPRG)
- 7 complaint, objection
- 8 retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

(Other possible fields include Nationality: e.g. U.S.; Definition: e.g. of Insider Trading, SALT-II; Time: e.g. future. None is commonly used.)

Fig. 1. Example TREC request (topic).

mutated, even honed, and have a complex structure with several distinct fields. The Narrative field, in particular, clearly indicates what properties documents must have in order to be deemed relevant, thus reflecting the user's needs; and the Concepts field contains a set of chosen, often compound terms, already constituting a 'proto-query'. At the same time, there are 'implicit' category terms like "US company" in Narratives that are; problem for indexing, and proper names that are a challenge for matching. Thus, the TREC requests look like standing interest, or routing, profiles, and were indeed designed with this task in mind; in the TREC tests they have, however, been used both for routing and for one-off or 'ad hoc' searches. Finally, as the foregoing suggests, the TREC requests were provided by professionals, experienced in request formulation and, further by users whose primary task is to do information analysis (i.e., not typical of many user communities).

With respect to relevance needs in TREC, one important point is that documents do not have to be wholly, or even primarily, about a request topic; a document may be relevant given only the facts conveyed in two independent sentences. The implication is also though this is not stated but follows from the routing background, that high recall is the normal requirement.

The relevance assessments are on the combined search output for all the participants official submitted runs. The official specification limits the number of runs per participant and sets a standard limit on

output size; at the same time, there are many participant! applying very different approaches; the *relevance pool* is therefore likely to be less strongly or arbitrarily biased than for many collections in the past and, since the various output! retrieved different documents, is likely to have pulled in many, even most, of the relevant documents. The number of documents assessed for relevance per request was large; the assessment was not done by the requester, but by similar professionals. (In the rest of the paper *relevant document* always refers to those so known from assessing the pool.)

Reviewing the TREC data from the point of view of *test collections*, the major points are as follows.

First, it is actually a *family* of related collections; the evaluation design described below has meant that tests have been done on various combinations of document and request subsets, with a document subset defined as one of the three blocks D1, D2, or D3, and wit a request subset one of the three subsets numbered 1-50, 51-100, or 101-150. The subsets have been set up over time, but though some sources are not represented in each document subset, and neither document nor request subsets have been drawn as systematic sample from prior wholes, they are similar enough, or at any rate sufficiently typical of collection growth, to be taken together. This is important, because whereas the document sets al large, the request subsets are small and only taken together form a fair set; so caution may be needed in drawing inferences from subset performance. For convenient reference in this paper, I shall refer to the whole body of material as the *TREC Collection*, and use *TREC collections* when the division into subsets is relevant.

Second, as a test collection, the TREC Collection is strikingly large, both absolutely and by comparison with those previously used for IR research. This refers not only to the number of documents, in particular, but also to the fact that they are largely full-text one: A test collection of this size constitutes a completely new test environment for the field.

Third, the Collection is a constructed rather than natural (or proper sample) one. This description applies in various ways. It applies to the aggregation of material from very different sources, not all of which might be searched directly with requests of the type used, but rather excluded via checks on database descriptions. It also applies to the requests These were specially developed for the TREC programme and were not drawn from ordinary retrieval operations, though they were intended to be like those used in practice. They were also made by only a small number of experts, so the number of requests does not represent the same, or even a large number, of users. Finally, the relevance assessments were not made by the request providers, and the data gathering design for these also meant that there is no reflection in the assessments of the number or choice of documents that actus users would have identified in practice. Of course, the important other requirements for relevance data for test collections that lead to pooling and obtaining as many assessed documents as possible are themselves an unavoidable source of artificiality; but the lack of reference to original requester's judgments makes the TREC Collection like former collection where proxies were used, even if the requests themselves provided unusually detailed guidance for the assessment.

The difficulties of forming truly 'real' test collections are well known, and relevance data in particular have to satisfy fundamentally conflicting demands. The points just mad are not intended as criticisms; the TREC data provision has been a heroic effort, thought fully grounded and carefully executed, and drawing on past IR test experience. My point here has only been to note salient features of the Collection possibly bearing on future conclusions from the tests. Moreover, although the TREC Collection is perhaps most accurately defined as semi-real rather than real, some of its properties, for instance the document heterogeneity, may be indicative of future search conditions and also constitute so taxing a test environment that good results obtained in it will carry over elsewhere. Over all, the TREC requests are the most critical element in the test data. The fact that they an rich in content, carefully designed, and elaborately structured; are intended for document routing; and come from information professionals clearly has implications for inference: from TREC. Thus, it is important to consider the extent to which performance results it the tests reflect generalised *tailoring* to the type of data found in TREC, notably the full text documents, or more specific *tuning* based on the particular properties of the test material, notably the requests.

Equally, although it is certainly helpful to have large relevance sets based on extensive pooling, it would be desirable to have some study of the nature and degree of bias there may be in the relevance sets.

2.3 Evaluation paradigm

The evaluation paradigm being used subsumes both what the IR task is taken to be, and hence what *assessment gauges* are used, and the test organisation. In terms of the analysis of system evaluation given in Galliers and Sparck Jones (1993), the TREC tests have a *remit* determining the rest goal, and a design specifying gauges, data, and procedures, with *remit* and *design* together constituting the evaluation *scenario*. The goal stems partly from the TREC motivation, but meshes with conventional approaches to IR testing done outside the real working environment of users and operational system. Thus, the goal is taken, strictly, just to be capturing relevant documents, which leads in a very natural way to the use of *recall (R)* and *precision (P)* as evaluation *criteria* with specific performance *measures* using output cutoffs and application *methods* for these measures averaging across values for individual requests. The measures give *P* at standard recall levels, and at standard document output size (rank) levels, from which single number average *P* and single number *P* for total relevant can be, respectively, derived. (Values for *fallout* are also computed, though these have not been accorded much attention.) The general approach is that common in the field, and popularised by the SMART Project, the different measures reflecting slightly different views of *R* and *P*; but document cutoff is better suited to intuitive interpretation and more realistic. *P* performance for very high *R* is not normally of practical interest. There are also well known problems in assessing iterative search performance; in TREC, where comparisons across participating teams cannot be tightly controlled, results have been computed using the frozen ranks method.

The test design has two distinctive and especially important features. The first is the way requests have been designated as *routing* or as *ad hoc*, and the closely related way in which request and document subsets have been combined for tests. The requests were, as mentioned earlier, designed for routing. Routing normally implies allocating documents from an incoming stream to members of a user community. But searches for these requests were in fact done on static document sets as wholes, thus changing the IR situation, with implications for performance conditions and assessment; the sense in which the requests were properly treated as routing ones was in the way relevance information for searches on earlier document sets provided *training sets* that could be used to develop and modify queries for application to new *test sets*. At the same time, the requests were put to double use, since they were also treated as *ad hoc*, for one-off searches of an existing document collection, though this was not the purpose for which they were designed or would normally be put. This is not to suggest that a single request may nor be legitimately put to very different search uses, but in assessing TREC it has to be borne in mind that the *ad hoc* requests might not be typical, because of being developed with the care and attention that routing requests can justify.

Thus, in the TREC evaluation, the same requests are used either in *ad hoc* or in routing *mode*, according to whether relevance information from *previous* document sets is available for building queries from them. This dual-purpose *modus operandi* is associated with the elaborate succession structure for tests and their definition in terms of individual test collections. In general, for some specific query set, this can be applied to some new document set in *ad hoc mode*, for *ad hoc* performance evaluation. This set, together with the relevance assessments made for the *ad hoc* test, can then be taken as training data for the same queries, for future application in *routing mode*. This routing application is therefore of old queries to a new document set, to which a new set of queries is also applied in *ad hoc mode*. The evaluation strategy for the TREC programme as a whole thus involves a kind of rolling procedure over related collections, with the two-mode operation, in a sense, enlarging the test collection.

The other important feature of the evaluation paradigm, at least as administered within the TREC programme and so constraining its participants, was the limit placed in the number of official *submitted results* per phase: no more than two routing runs and two *ad hoc* runs per team. This has obvious and immediate advantages for comparative purposes and generally telling the wood from the trees. It has, of course, not stopped participants doing and publishing other runs, and the submitted results should

emphatically not be taken as all there is to say about the IR approaches they represent. But it is important to recognise that as TREC is a multi-party enterprise, its evaluation paradigm has this wider manifestation relevant to assessing its results. (Other aspects of the TREC programme methodology, like the questionnaires, that bear on research practice rather than evaluation per se are considered later.)

2.4 Generic context

The TREC programme and enterprise has naturally, like all the (D) ARPA evaluations, evolved a house style. This in part follows from its motivation and evaluation paradigm, but there is also what may be called its *generic retrieval system context* to take into account, as there are constraints or presuppositions associated with this that may influence the approaches adopted by the participating teams. This *generic context* is not only, and obviously, that indexing and retrieval are essentially, even if not wholly, automatic, but also that users, when receiving matches for routing requests or engaged with adhoc ones, will be interacting with a computational system. Though the nature of the interface itself has not been considered at all, and tests with humans doing interactive searches have so far been very limited, there is an underlying presupposition that computational interaction is the norm. This may indeed cover the entire use of the full text, and not just preliminary or superficial scanning. (Both the interactive tests and also the relevance assessment have required interfaces with some of the properties that would be needed for end users in normal circumstances, but this has not been the prime design consideration in either case.)

The computational context in itself implies both opportunities and restrictions for indexing and retrieval. But the size of the document set, the fact that in the real case it would grow and change, perhaps quite drastically, and the provision of full text, jointly define a generic retrieval context not extensively analysed in evaluations to date, but having important consequences. While these do not always apply in a strictly routing situation, they can; and as TREC itself, and others wishing to exploit its findings are nor so confined, the computational features of the TREC case deserve notice. They imply:

1. it is foolish to rely primarily on any kind of relatively fixed, controlled, indexing language;
2. but it is also dangerous to assume that any indexing vocabulary, like a set of term classes, depending on properties of the document set as a whole can be reformulated at intervals;
3. even if it is feasible, it is not going to be worth applying elaborate (pre)processing to each document designed to provide a full and complete index description, especially as an autonomous surrogate replacing the source text;
4. at any rate, it is not sensible to assume that elaborate initial descriptions can be reformulated, for entire files, to suit changing document or request file conditions;
5. it will be too hard, or too costly, to construct sufficiently powerful explicit Boolean queries for effective retrieval (apart, say, from long-standing routing profiles); 6. but it is essential to supply resources and mechanisms for request formulation and query development.

The consequences of these points taken together are, first, for indexing and retrieval philosophy and, second, for operational system design. Thus, with respect to retrieval philosophy, there is a presumption that natural language is straightforwardly the language of indexing; that indexing is request-based (documents are indexed via their query matches); that output ranking allows discrimination and flexibility; and, therefore, that the focus of attack in retrieval should be on the development or modification of source requests and the actual search queries derived from them. The prime questions are, thus, what form of natural language indexing, and of query manipulation, is either generally optimal or best for given conditions.

From the operational point of view, the issues are, first, what forms of information have to be gathered at document file time and what implications search uses have for how this information is stored and accessed; and, second, what information resources, and display facilities, query development requires.

While this generic context seems to be largely taken for granted by TREC participants, the TREC effort has included some explicit investigations of practical questions relevant, to document file handling; and

the TREC enterprise as a whole can be seen as implicitly testing all the presumptions just listed by showing that although they may imply a somewhat crude or limited approach to indexing and retrieval, this can still deliver reasonable performance, even if it may not be the best. The retrieval style just described also supplies a natural *baseline* for performance in the form of single terms (word stems), with simple weighting using inverse document frequency and perhaps also within-document frequency, and sum or product scoring for matches. The immediate questions, then, to be answered for a TREC-specific baseline are whether within-query weights are also to be used, and what parts (fields) in the request.

Given this baseline, the challenge is, thus, to improve on it by moving in any or all of the following directions: using more complex terms; being more selective on terms; exploiting other or additional weighting; utilising supplementary vocabulary resources or query expansion. These possibilities apply to both routing and adhoc requests, from a static perspective. There is also the dynamic perspective allowing learning, referring on the one hand to system *adaptation* over time and applying to all requests (e.g., through changes in term associations), and on the other to *feedback* bearing on individual requests. This may be over time in the routing case or through search iteration in the adhoc case, and normally exploits relevance information in *relevance feedback*. The baseline case here may be deemed to cover only the automatic changes in inverse document frequency weights that naturally follow collection changes, and *minimal feedback* relying just on user document assessments to reweight query terms. The corresponding lines to pursue for performance improvement therefore cover, first, elaborations of this basic reweighting and, second, combinations of learning with the other possibilities just mentioned referring to complex terms, term selection and weighting criteria, and supplementary resources; they thus lead to dynamic term classification, for instance and, most importantly, to query alteration by expansion or deletion. This has a whole range of options with more or less decision by the system or user.

Finally, the generic context for TREC makes technology considerations worthy of study in their own right. Though the test environment imposes its own data-organisation requirements, for instance, high modularity to allow system alternatives, it is essential to investigate data management issues, like file structures, paging, etc., because the data bulk imposes operational constraints, and the choices for data organisation interact with indexing and retrieval ones to affect not only efficiency, but effectiveness.

2.5 TREC effort

The final important feature of TREC, in relation to IR R&D, is the sheer scale of the evaluation effort in terms of the number of teams involved and the consequent range of approaches being tested. Since the programme in itself explicitly involves a common treatment of output for common performance measures, the provision of replies to a detailed procedure questionnaire, and workshops at the end of each phase, there is a very rapid dissemination of results to, and interaction within, the participating community, leading to a quick response to results and also, via the active programme committee, to continual development of the whole assessment strategy. From this point of view, TREC is not merely a large effort, but a *de facto* cooperation as much as competition, even if the cooperation is indirect through changes in individual approaches rather than through explicit joint work. This has, however, also occurred, and there is a further, important direct form of community cooperation in the work of the programme committee. All of this has meant that a very large amount of ground has been covered, and much more quickly than in previous IR work, with some consequences for assessing the TREC findings.

In the next section I shall consider the actual work done by the TREC participants, and in the following one summarise their test results; then in assessing the results, I shall try to establish what influence the TREC case had on these and, hence, on the conclusions to be drawn about needs for further research and for the wider exploitation of the results in operational systems.

3. TREC WORK

The participants in any TREC phase are divided into two categories, with Category B participants in 'startup' mode doing smaller-scale tests. Because in reporting the TREC results in the next section I shall focus on the full official tests, I shall not consider Category B work here, other than where it is especially appropriate. I shall also, in this overall review, not single out individual teams for concentrated attention; however, as references to teams are needed, I shall cite these-by simple institution names, for which a key is provided in Appendix A. Also, unless specific citations are made to other publications, the references for teams' work are to their contributions to the TREC-2 Proceedings (Harman, 1994a).

3.1 Performance factors

As noted, the TREC(-2) participants have pursued a whole range of approaches to indexing and retrieval. It is helpful, in reviewing the TREC tests and the lessons to be learnt from them about the different approaches, to characterise these in relation to the evaluation framework elaborated in Galliers and Sparck Jones (1993) and mentioned earlier, where *performance factors* are distinguished either as *environment variables* with *values*, or as *system parameters* with *settings*.^{*}

3.1.1 Environment variables

In TREC the environment for all the participants in any one phase (and category) is the same, being the documents, requests, request mode (routing or *ad hoc*), relevance needs, and relevance data for previous document sets. Thus, for each phase there are two environments defined, respectively, by mode, each with their test collection defined by request set (with relevance needs), document set and (subsequent) assessment set, and also prior test collection now available as training collection. However, while the phase environments for a mode are technically distinct, they do not, for the reasons mentioned earlier, cover any serious variation in the values of the environment variables; the successive test collections are designed to be like one another. The TREC tests do not therefore involve the significant changes of environment properly required to check on the behaviour of system parameters and fill out the test grid cells. Moreover, while the mode environments are distinguished in particular by the supply of relevance information, the fact that the same requests are used, albeit in different set combinations, makes the environments less different than appears. Thus, whereas the main thrust of the TREC tests over time may be taken as intended to establish that systems perform consistently across test collections (setting aside changes to the systems themselves), the collections have not been varied enough to demonstrate this or, as research understanding requires, explain why they perform as they do. It is, thus, legitimate to characterise TREC, at a coarser level, as having a single, constant environment, and the TREC work so far as not involving any serious attempts to explore the effects of environment changes with, for example, very different types of requests.

3.1.2 System parameters

The real variation in the TREC tests has come from the differences in system parameter *settings*, both at major and minor levels across and within the different teams' tests. Here TREC has filled in many test grid slots, though these are somewhat unevenly distributed and are largely arbitrarily related to one another, especially at any one level of detail, with many gaps. The TREC evaluation thus addresses the question: What differences do these variations make, in fact, at several levels of granularity, both coarse- and fine-grained, where coarse-grained may be illustrated by the contrast between linguistically and statistically defined compound terms and fine-grained by choices of constants in weighting formulae.

^{*}"Variable" and "parameter" were used the other way round in my earlier papers on IR tests, especially reports of the 1970s, the present use fits better with other conventions.

Further, as noted earlier, the TREC research enterprise also applies a fixed evaluation context for all participants as defined by the standard evaluation scenario used: so while *R* and *P* are natural first performance criteria, there has not been any system comparison on other effectiveness criteria, and only very limited studies of efficiency.

Although it is not appropriate to describe individual TREC approaches and results in detail, a broad categorisation of approaches is needed both to assess the scope of the TREC, studies and draw conclusions from the results. The categorisation refers primarily to matters like the type of treatment given to documents, which are clearly performance factors. But it is also useful, for a comprehensive view, to extend the categorisation to include the general form of model underpinning an approach, for example, whether an overtly linguistic or statistical view of information is being taken. Thus, the choice of *indexing and retrieval model* can be treated as defined by a highest-level parameter under which lower-level parameters fall into groups concerned with the nature of the *indexing vocabulary*; of the *indexing strategy* for documents (i.e., the nature of *document descriptions*); of the *document indexing sources* for these descriptions; and, since documents and requests may be treated differently, the nature of request descriptions (i.e., of *queries* and *query sources*, including not only requests, but also, e.g., relevant documents). There are, further, the *search strategy* and *scoring criteria*, clearly separate because the search procedure may affect what documents are inspected for matching; and *output form*.

In some cases there is a strict dependence between settings for these different parameter groups, in other cases natural or habitual associations, in yet others complete independence. These heterogeneous relations between performance factors as represented across TREC approaches make it very difficult to assess the general implications of the results for particular approaches or even types of approach, because it may be hard to attribute responsibility for these results to any particular features (i.e., choices of setting for system parameters) of the approaches involved. This is well illustrated, for instance, by Buckley's discussion of weighting (Buckley, 1993).

Moreover, the classes of performance factor listed earlier are themselves informal, though familiar and to some extent hallowed by tradition. Thus, in practice, whether one is dealing with indexing vocabulary or document description may be hard to determine, and it is in any case necessary to distinguish logical characteristics from practical implementations. However, the categories allow a first cut across the range of approaches with a view to attributing performance findings to system features.

It is also necessary to recognise *learning* as a high-level system feature. This may be applied in relation to various system elements, but is conveniently treated under its own heading.

In discussing the different approaches adopted in TREC for each parameter, in the material that follows I shall put parameter questions; these are intended to focus the comparisons between results in the next section, where I shall try to answer them.

3.2 Indexing and retrieval models

Characterising the models underlying the TREC approaches is not easy. But it may be done in terms of what criteria are used to choose index terms (i.e., how they are grounded as units, irrespective of whether these are in fact single or multiword units). From this point of view, units consisting either of single words or of multiple words held together only by proximity-based co-occurrence can be labelled statistically motivated, whereas term units defined using explicit linguistic (syntactic or semantic) constraints (e.g., selecting head nouns) are linguistic. These alternatives may be instantiated either through prior vocabulary definition or at file or search time. By extension, word classes may be (explicitly) statistically or linguistically motivated. This statistical/linguistic distinction is one informing an entire indexing and searching process; however, some TREC teams apply their model in a much more aggressive and all-encompassing way than others. In many cases, too, underlying models are diluted, overlaid, or mixed through complex combinations of resources and procedures. The categorisation made here is thus only indicative.

The statistical models represented in TREC include the vector model associated with the SMART system (e.g., Cornell, TMC) and the 'straight' probabilistic one (e.g., City, Dortmund); Latent semantic indexing (Bellcore) is a type of vector model, probabilistic inference nets (e.g., Amherst, CUNY) are another form of probabilistic model, which also extends towards learning (see below) via connectionist implementations.

These models ramify into all aspects of indexing and retrieval. The TREC linguistically motivated models are less aggressive, and are primarily focused on the use of syntactic or semantic criteria for identifying index terms, especially compound ones, and variant representations of the same concept (e.g., CMU, NYU). Selection criteria based on large-scale discourse structure have also been examined, by Syracuse. In general in TREC, however, there has been little use of linguistically complex complete index descriptions of the traditional subject-heading type, presumably both because the TREC requests are too extensive to encapsulate in one or two descriptive units, and because this is not appropriate for the documents. Some projects (e.g., Siemens, Conquest) have also used manual thesauri or other classifications implicitly based on linguistic considerations. In general, linguistic models for index terms have been combined with subsequent statistically motivated processing (e.g., for weighting or scoring). Equally, primarily statistical indexing may be softened by encompassing compound or phrasal terms based on adjacency or proximity (e.g., Dortmund, TRW). TREC therefore reflects the fact that the value of linguistic models has been a longstanding issue for IR (cf. Lewis & Sparck Jones, 1993).

Model questions. The first questions to be asked about the TREC rests are thus what differences alternative models make, as follows:

- M1: Are linguistically motivated models superior to statistically motivated ones?
- M2: Are there performance differences between the models in either class, for instance, between vector and probabilistic in the statistical group, between sentence syntax (NYU) and discourse syntax (CMU, weakly, Syracuse) in the linguistic?
- M3: From a different perspective, do more sophisticated or refined models contribute anything, where linguistically motivated term selection criteria may be taken as more refined than statistical ones; or in the Linguistic case, do richer linguistic models (NYU, Syracuse) outperform plain ones (CMU); or in the statistical case, are rigorous probabilistic models (City) superior to vector ones (Cornell), or fancy vectors (Bellcore, HNC) better than simple ones (Cornell)?
- M4: As the main specific issue here (given that everyone does stemming, and starting from the baseline): Are linguistically grounded compound terms, as opposed to term conjunction in matching, valuable?

3.3 Indexing vocabulary

The model distinctions just considered clearly apply to the indexing vocabulary used. The issue for vocabulary is whether a vocabulary is established as a (more or less) fixed entity for explicitly characterising documents, rather than being simply taken as the universe of potential terms embodied in the documents and realised at search time. As mentioned already, operating with an explicit vocabulary is so exigent that it has not been common in the TREC work, though it is exemplified in the statistical case by latent semantic indexing (Bellcore), and in the use of collection-motivated phrase vocabularies by Cornell and Dortmund. (Manual thesauri have been primarily used to supplement, not to form, descriptions, e.g., by Verity, Siemens, Conquest.) However, vocabulary design is also approximated by term weighting schemes using collection rather than individual document information, since these can be taken as relativising term value for any use. In this form, vocabulary manipulation is common to both linguistic and statistical approaches and for the latter to, for instance, both vector and probabilistic cases, as well as being a motivation for the learning strategies considered later.

Vocabulary questions.

- V1: Does a holistic approach to the indexing vocabulary pay its rent?
- V2: Is linguistic sophistication important?

However, it should be noted that the TREC requests (and also documents) include proper name elements with many variations, abbreviations, etc., and several teams have made use of special-purpose lexicons, grammars, or other processing strategies to handle these terms. Whether these are mandatory because matches on these items are demanded, or just assist in generally improving performance, depends on careful analysis of the individual application. The claim is that they are beneficial, though not hugely so, in TREC.

3.4 Document descriptions

The model distinctions clearly apply to the way individual documents are treated (i.e. what the basis is for choosing a term to index a given document). Thus, setting aside the virtually universal use of stemming and a stoplist, the main distinction between the TREC teams has been whether document terms are linguistically or statistically motivated, the latter including proximity, which is also *de facto* document indexing, even if implemented through matching conditions at search time. The context vectors used by HNC (cf. Gallant 1991) are a complex way of exploiting proximity information in statistical document indexing. Linguistic processing is mostly local, sentence-based (e.g., Amherst, NYU) but, as mentioned earlier, may take account of large-scale document structure, though the Syracuse approach here has not been actually tested in retrieval yet. Using compound terms (whether linguistically or statistically defined) or single terms is itself a major distinction, and wherever compound terms are used, one of the issues is whether their component elements are also assigned as terms or allowed in matching. At the same time, many approaches exploit statistical weighting for whatever document terms are used, with the practical problem for derived compounds defined at search time of having the incidence data to do this.

Description questions.

- D1: Is linguistically motivated indexing superior to statistically motivated indexing, especially where this is derivative for individual documents?
- D2: Are compound terms (whether linguistic or statistical) superior to single terms?
- D3: For the full texts used in TREC, what is the value of document-specific weighting schemes?

3.5 Indexing sources

Whereas indexing strategies may be indifferently applied to documents or requests, there may be document-specific constraints, depending either on views about document form or on the practicalities of large file searching. As mentioned earlier, a choice that might be thought relevant to the full-text case, namely, whether full text or only title+abstract is actually used for searching, is not often available in the TREC case; and for the same reason, differential weighting taking advantage of the concentration of key information in abstracts is not possible.

Some strategies involving *subdocuments*, for example, paragraphs (Cornell) or pages (CITRI), have, however, been investigated within TREC that can be looked at from this point of view. Thus, preferring (Cornell, CMU), or requiring (CITRI, CUNY) query matches within page units can be viewed as an opportunistic use of subdocuments as indexing sources for the whole. However, retrieving, that is, offering the user, subdocuments only has hardly been investigated in TREC, since though the nature of the TREC requests might suggest it could be appropriate for long file documents, even if not worth it or suited to the short news stories, the TREC relevance data makes proper performance evaluation impossible (CITRI).

Source questions.

- S1: Is using subdocuments as indexing sources, as opposed to the whole text, useful?

3.6 Queries and query sources

Although IR of course requires compatible document and request indexing, and some TREC

participants adopt similar strategies for both, most of the TREC effort has gone into request rather than document indexing. The complex and non-standard character of the TREC requests means that they offer varied possibilities for query formation, and work has been done both on different ways of treating the various fields (e.g., Amherst, Syracuse) and on which fields are useful. The difficulties of processing the large document files has also encouraged concentration on the requests, where manual query formation may be viable or manual modification perfectly practicable (e.g., Rutgers, TRW, Verity, Conquest, PRC). Of course, queries as search specifications may be very complex (e.g., Amherst, CMU, TRW, Verity, Conquest), or have a (semi) boolean form (e.g., GE, TRW, CUNY, VT), though actual document matches are simple. Compound index terms defined by proximity may be specified via queries (e.g., GE, TRW, VT), and may be treated slightly differently in queries and documents, for instance, where compound terms are only accepted as such on tight criteria for requests, even if they are more loosely matched in searching; though formally the indexing is the same, the grounding is biased to the request. The same weighting schemes may be applied to documents and queries, but though some types may be less informative, for the latter the long TREC requests offer scope for query term weighting, and this has been extensively used (e.g., Cornell, Bellcore, ETH, Verity Conquest). Apart from the manual queries (which may be linguistically complex), query indexing has been primarily statistical, though some teams have applied linguistic analysis to request texts (Amherst, Syracuse, NYU, CMU).

Further possibilities specifically for queries are associated with mechanisms for *query expansion*, and especially query expansion in relevance feedback. Many teams used relevance feedback to apply training data about documents including relevance information to alter individual queries (as learning this is considered later). This is natural for routing requests, but there have also been a few tests with adhoc requests (City, HNC). Several teams used thesaurus expansion in conventional style (e.g., GE, TRW, Verity, Siemens, Conquest), without relevance information; this might or might not be automatic, and use a statistical or linguistic thesaurus. However, TREC offered the opportunity for a deliberate and systematic use of relevance information for query expansion, not just reweighting, and as a device for improving either routing or adhoc queries. The expansion process may be fully automatic (given relevance assessments), semi-automatic (the user may be offered vetted terms to endorse), or manual. Many TREC teams used expansion for the routing mode (e.g., Cornell, Dortmund, Amherst, HNC, Bellcore, CMU), but experimental difficulties limited work on the adhoc case (City).

With respect to query sources in requests, the nature of the TREC requests made this quite a complex matter. The request fields differ in type of information, degree of formality, and linguistic properties (isolated words or phrases versus running text), though they overlap lexically. The range of fields available in TREC stimulated many tests trying various selections from and combinations of fields, with a specific point of interest in the utility of the Narrative field (Amherst, VT). Further, as both requests as wholes and some individual fields are quite long, this has implications for query length and query term weighting as system parameters.

Query questions. Thus, with respect to queries, there are both questions that are relatively independent of the particular TREC application, and ones that are more narrowly tied to it.

- Q1: What form of index description is effective, especially in complex superior to simple?
- Q2: Is manual query formation better than automatic?
- Q3: Is query expansion valuable?
- Q4: Is feedback in general useful?
- Q5: For TREC specifically, are explicit statements of document relevance criteria helpful as a direct source of search terms?

3.7 Search strategy

The search strategy in its fullest sense has not been a parameter of major interest in TREC, other than via indexing or scoring procedures applied in a routine and uniform manner across all queries at search time. More generally, the TREC evaluation design working with ranked output implies a 'normal' strategy of ordering by decreasing matching score. But there has been almost no study of classic areas

of concern like, for example, starting narrow and broadening out, or vice versa, except as an automatic filtering device, or (except for Cornell) of flexible and varied procedures suited to the individual request; everything is standardised and averaged. Although some queries were manually constructed, this was not by end users, and apparently with little study of or control over the factors influencing the construction process. Search strategies are also applied in the reformulation of queries in iterative searching (as in the City case), but tests in this area have been too limited for explicit strategy studies, and indeed the motivation for these tests was such that those developing searches were deliberately encouraged to do whatever they thought was sensible. Moreover, much of the motivation in feedback tests has been to develop automated methods that reduce effort for the end user. Using subdocuments as a normal target in matching (CITRI, CUNY) can be seen as a type of search strategy; but it has been uniformly applied without variation for individual requests.

The very detailed and specific nature of the TREC requests has, in fact, reduced the pressure to devise productive ways of developing useful queries from tentative and minimal initial requests, or to tailor searches according to types of request, with one strategy for a broad request and another for a narrow one. The TREC enterprise has thus not yet addressed a major challenge for operational systems faced with very large volumes of material and, also, not very experienced users.

Search strategy questions. There is a generally relevant question about searching that can be asked of the TREC tests:

- Y1: Do uniform and primarily automatic strategies give reasonable results for one off requests, and also reasonably uniform results across different queries?

Though the TREC requests do not vary greatly in specificity, they still differ enough for it to be legitimate to ask these questions.

3.8 Scoring criteria

Scoring criteria, on the other hand, have been investigated in various ways. While matching scores may be strongly influenced by the form of indexing (e.g., by document derived term weights, there are nevertheless different possibilities for characterising the relation between query and document. Thus, the SMART model, used by several TREC teams, embodies a specific definition of what a document-query match means.

However, it is also possible to refine on query-document scoring by such strategies as promoting documents that match locally rather than just globally (Cornell, CMU). Some of the TREC participants have, moreover, experimented with *data fusion* ideas, or ways of combining different versions or subcomponents of queries (e.g., Rutgers, Amherst, HNC, IDS, VT, PRC). These naturally follow from, for instance, inference net models that can subsume different types of indexing information, but it is additionally possible to assign different scoring values to different information types as a function of query document matches, as investigated by, for instance, VT.

Scoring questions.

- C1: Are complex, differential scoring functions superior to more straightforward ones of a plain sum or product kind?

3.9 Output form

Output form is an important parameter for operational systems, covering both the way the search output for a query is presented (e.g., in rank order) and the way the user is given information about individual documents. Some of the issues arising in older systems where there was no output ranking, for instance, or only title information in the file do not apply in TREC. However, though full-text systems present their own challenges for output presentation, these have not been investigated in TREC (other than for page output by CITRI), since the provision of output just as ranked document identifiers is geared

wholly to the formal evaluation protocol and is divorced from users. The interface provided for gathering relevance assessments, though it might have properties bearing on output forms for users, was designed specifically for the evaluation task, and so should not be considered under this parameter heading.

3.10 Learning

I have explicitly or implicitly referred to learning under several of the previous headings. Whether or not a system is designed to learn at all is in itself a very generic feature, rather than a system parameter: the different points to which learning is applied, whether of data or process, are those that define parameters. However, learning may be more wide-spread, in system *adaptation* affecting all index term weights and relationships; or targetted on a single component, as in *revision* of the indexing vocabulary; or particularised by events as in individual query *alteration*, subsuming query *feedback* either as routing query *evolution* or as adhoc query *modification*, using *relevance feedback* for reweighting or expansion. These possibilities have all been studied before, but both the TREC data and the structure of the TREC programme offer a valuable context for research on system learning in response to changes in documents, requests, or relevance needs. First, TREC offers very large volumes of text data, providing a respectable corpus for learning word collocation patterns, for instance (e.g., GE); second, there are large relevance sets per request; and third, there is the routing task, for which query evolution is naturally appropriate. In addition, TREC has not depended on the use of a conventional controlled vocabulary or classification of the sort that is manually revised either via adhoc gap plugging or as an occasional, global exercise that relies on informed experience but is not rigorously tied to collection data. With TREC it is in principle possible to investigate continuous and systematic learning, and also to take advantage of the modern technology and learning models (e.g., connectionist ones) that are available. Thus, the TREC effort has included work on automatic learning with a fairly serious intention, including overall system index adaptation (CUNY), adaptation centred on vocabulary revision (Bellcore), and extensive work with relevance feedback by many teams. But the adaptation work has not been pushed far as a continuous rather than spasmodic process, and the relevance feedback studies, even though they have been notionally for routing, have also been noncontinuous.

Adaptation is of course collection *tuning*, whether within the context of a standard IR strategy or one already *tailored* to a collection type. One important form of tuning for which TREC, through the volume of data it provides, has offered a serious study opportunity is via the application of *regression analysis* to determine the relative weighting to be given to the various elements of indexing (Berkeley, Dortmund, ADS).

Learning questions.

- L1: Is overall adaptive tuning to a collection valuable?
- L1: Is refined vocabulary revision helpful?
- L3: With respect to requests, does relevance feedback significantly improve performance?
- L4: Does regression analysis help tuning?

4. TREC RESULTS

In this section, I shall indicate what, in my view, the TREC *results* are (i.e., what the performance data laid out in the evaluation record, and especially, though not exclusively, in the official submitted results, show). This section is thus primarily a descriptive summary that tries to pull together the very large amount of very varied performance data, especially, as mentioned earlier, for the full tests as opposed to Category B ones. I shall then as a separate exercise, in the following section, review these results, taking into account the participants' own accounts of their aims and activities, in an attempt to determine what the results mean and, hence, to establish what the TREC *findings* about IR systems are. This broader assessment has naturally to refer to the influence of the TREC case.

4.1 Review strategy

My summary of the TREC results has necessarily to deal in rather crude generalisations. The complex test context, wide spread of runs, and hence lack of systematic, finegrained experimental control make this the only practicable review strategy. It is also important, given the many questions still open, not to attempt a vulgar 'pick the winners' approach.

In comparing performance figures, it is impossible to be anything other than informal. Though systematic statistical significance tests are needed, they have not yet been done, and it is also not clear how the magnitude of performance differences for this data are to be interpreted. For example, to refer to an earlier scheme, is it reasonable to label performance differences (assumed statistically significant) of 5% and 10% as 'noticeable' and 'material', respectively? These conventions also referred to *R/P* graph areas. In this paper I shall, in my attempt to make a first cut across the result, further focus my comparisons on one *fixed point*, namely, *P* at document cutoff level 30, justifying this on the grounds that this is both a transparent and a realistic performance indicator, at least more so than the others used. I shall then, in comparing two performance results at this point, use the notions just mentioned, namely, no difference, a noticeable, and a material difference (symbolised as =, >, or >>), with the last two corresponding (roughly) to 10% and 20% increases in *P*; this is appropriate for such simple analysis.

It is, further, useful to summarise the submitted full results for the fixed point *Pdoc30*, as shown in Table 1. This assigns each team's best result for each of their adhoc and routing pairs to a performance *block*, and also refers only to the upper performance levels, namely for adhoc the blocks $\geq$ 55, 50, 45, and for routing $\geq$ 60, 55, 50. (For the comparative record, analogous results for recall cutoff at 30% are also shown.) Results or approaches where there are no or few full runs are considered only where this is especially relevant; it is not really appropriate, in particular, to discuss performance results for TREC starters who are still getting their act together. It should be noted that, quite apart from any strategy variations that naturally follow from the difference between routing and adhoc retrieval, the results for adhoc and routing for a given team are not necessarily obtained in a similar way; although the general style is normally the same, teams may have investigated distinct approaches in the two cases.

The table, of course, somewhat crudely divorces adjoining figures (e.g., two teams with results of 60 and 59, respectively, are separated, so the table blocks should be treated with caution and only as a way of achieving leverage and insight). It should also be noted that the recall cutoff values are somewhat lower than the document cutoff ones, indicating the danger of attributing too much importance to the absolute numbers obtained, though the relative positions of teams according to the two measures are similar.

It also cannot be emphasised too strongly that these figures provide only one cut across the large mass of experiments and that the detailed accounts, especially for the final TREC-2 Proceedings, make it plain both that sheer accidents can occur (e.g., Bellcore, Cornell, probably or actually artificially depressing performance, and also that fuller work can significantly upgrade performance, for instance, in choosing suitable values for constants in formulae. Thus, work at City following the official runs submission raised the adhoc performance shown in Table 1 by two blocks, and or CUNY brought routing performance up to the Table levels. These two examples show how the continuing search for performance improvement can be one of the major gains from TREC and similar enterprises, while reinforcing my point that the officially tabulated runs are only a starting point for assessing TREC.

Table 1 Summary of better official results for full runs

Adhoc		Routing	
Precision at cutoff		Precision at cutoff	
Recall 30	DOCUMENTS 30	Recall 30	DOCUMENTS 30
>= 45 Dortmund Amherst CMU Siemens VT	>= 55 Amherst HNC VT	>= 55 Cornell >= 50 Dortmund	>= 60 Cornell Dortmund
>= 40 Cornell Berkeley HNC Bellcore CITRI CUNY	>= 50 Cornell Berkeley Dortmund CMU Verity Siemens CUNY	>= 45 City Berkeley Amherst Bellcore CMU GE TRW	>= 55 City Berkeley Amherst Bellcore CMU
>= 35 City ETH Verity Conquest	>= 45 City Bellcore ETH CITRI Conquest		>= 50 Rutgers HNC GE TRW Verity Siemens

Notes: Figures not rounded; teams per block in Proceedings page order; best of pairs if two runs submitted, but two often similar.

4.2 Global comments

Looking at Table 1, the following global comments can be made:

1. Performance is pretty good: 55% on hundreds of thousands of documents is very creditable.
2. The routing results are generally better, but not enormously so.
3. The spread from top to bottom even of this subset of better results is quite large.
4. Many teams have very similar performance.
5. Teams generally have similar relative positions for routing and for adhoc.

4.3 Question answers

Now, considering the test results in relation to the TREC participants' approaches to indexing and retrieval, I shall first seek answers to the system parameter questions asked in the previous section, and then sum up the major conclusions that can be drawn from the results overall. The question answers attempt to isolate performance elements but, as noted earlier, this can only be done in a rough and ready way, given the mass of miscellaneous combinations represented and the difficulty of assigning performance responsibility; equally, although individual parameter choices may not seem to contribute much, they may be effective in (some) combinations.

4.3.1 Model answers.

- M1: Heavily linguistic approaches are only represented by CMU, since NYU is in category B. CMU's performance shows that a linguistic approach performs perfectly well, but no better than statistical ones.
- M2: Different linguistic models were not assessed. The main statistical schools, vector and probabilistic, perform similarly (e.g., Cornell vs Amherst).
- M3: (linguistic not assessed.) Elaborate statistical models (e.g., Bellcore) do no better than simple ones (disregarding training effects).
- M4: Thus, following the previous answers, there is no gain from using linguistically motivated, as opposed to adjacency defined compound terms.

4.3.2 Vocabulary answers

- V1: Only one of the TREC teams applied any strongly holistic vocabulary design methods, namely Bellcore, but without obvious benefit. However, some (e.g., Cornell and Dortmund) employed a pruned and normalised extracted phrase list (along with ordinary single terms), though this too, alone, does not appear to be significantly superior to simple single-term extraction and individual-item weighting.
- V2: Manual thesauri were used primarily as adjuncts, but without noticeable effect (e.g., Siemens).

4.3.3 Document description questions

- D1: There is no clear gain from linguistic processing to select document terms (e.g., compounds, as in CMU).
- D2: There may, nevertheless, be a slight advantage from compounds (e.g., Amherst, Dortmund) as opposed to single terms.
- D3: There does appear, moreover, to be advantage in document-based weighting, practised by many participants, including, for example, Cornell, Dortmund.

4.3.4 Indexing source questions

- S1: Work using subdocuments to constrain matching as a basis for retrieval, as in Cornell, CITRI, is inconclusive.

4.3.5 Query and query source questions

As mentioned, the scale and structure of the TREC requests offered considerable scope for investigations of query formation, and many participants explored the possibilities. Thus, with respect to query formation, both manual and automatic approaches were tried (e.g., Verity, VT vs Cornell); also complex automatic strategies using statistics (e.g., Dortmund) or rules (e.g., CUNY) or linguistic processing (e.g., Amherst, CMU); query weighting (e.g., Dortmund, Bellcore); and various forms of query development through expansion and relevance feedback for reweighting or expansion. All of these last were particularly popular.

- Q1: There does not seem to be any striking advantage in applying sophisticated query construction methods (as such, independent of collection properties), whether relying on human intelligence or elaborate automatic strategies.
- Q2: There is no clear gain from manual query preparation.
- Q3: There does not appear to be any advantage in query expansion independent of relevance information (e.g., Siemens, Conquest).
- Q4: There does, however, appear to be utility in relevance feedback, especially via query expansion (primarily tested for the routing case).

With respect to query source (specifically for TREC), the Concepts field (cf. Fig. 1) is normally recognised as crucial, and Title and Description may be thrown in.

- Q5: Whether or not the Narrative field, giving relevance constraints, should be exploited, as a source of terms, term conditions, or weighting, is not clear.

The general implications of these answers, given the unusual character of the TREC requests to which they refer, are considered later.

4.3.6 Search strategy answers

Both individual teams, and different teams, obtained quite different performance per request (cf. Harman, 1994a).

Y1: However, uniform strategies applied across request sets gave aggregated results that were 'good enough'.

4.3.7 Scoring criteria answers

A few teams explored combination scoring or differential criteria applied at search time, under the headings of multiple queries or data fusion (e.g., Rutgers, HNC, VT). Differential scoring of a more pervasive sort may also be established by learning, for instance, through regression analysis as investigated by, for instance, Dortmund, and by choices of values for constants in formulae (e.g., City).

C1: There is not enough evidence to show that query-based differential scoring is valuable, though collection-based learnt differences may be useful.

4.3.8 Learning answers

The TREC experiments addressed learning primarily in two forms: generalised collection adaptation, or tuning, notably by regression, and individual query alteration in relevance feedback. Though vocabularies were constructed for TREC data sets, for example, by Bellcore, they were not continuously revised to match collection development, and though connectionist techniques were used by, for example, CUNY, HNC, they were not applied in a continuous learning mode.

L1: There is no strong evidence for adaptive tuning beyond the obvious weight changing that accompanies collection growth and decay.

L2: There is no proper evidence for the value of vocabulary revision.

L3: Use of relevance information for changing routing queries appears beneficial, though the TREC design limited the type of testing available. The TREC design restricted feedback testing too much for proper tests with feedback on adhoc queries.

L4: There is no clear evidence for regression analysis as a specific tuning technique, since systems without it worked as well as those with.

4.4 Summary of results

Summarising the results overall, there are no parameter choices that hold across all best performance: there are only tendencies for some system features to be associated with better performance, even if not required for it. However, combination effects may be equally important, given how much variation there also is between performances for approaches of similar types.

Subject to these cautions, we can now pull together the answers to the particular questions to make some broad brush generalisations across the results as a whole. These are

1. that very different approaches perform equally well;
2. that relatively simple statistical approaches along the lines successfully followed in past IR research perform as well as more complex ones (e.g., exploiting linguistic processing) in the 'long-document, large-document-set' case;
3. that insofar as compound terms are useful, which is not wholly certain, it may be sufficient to define these only by adjacency or proximity;
4. that the systematic exploitation of weighting, applied to different sources of information within the overall statistical framework, is valuable;
5. that tailoring to generic collection properties (e.g., long documents) may be useful, but this is not very clear;
6. that tuning to individual collection characteristics (e.g., through regression analysis) is desirable since though weighting in particular is useful, it is subject to considerable variation in value (cf. Buckley, 1993); however, such specialised aids as, for instance, name lists may not be helpful enough to demand explicit provision;

7. that concentrating effort on the request is much more effective and efficient than working (a priori) on document descriptions;
8. that query modification through expansion and reweighting is valuable.

As an alternative view, we can characterise the best performance compared with the baseline in terms of our earlier noticeable/material terms for improvement. For the baseline we can make use of TMC. This is not improper, as TMC explicitly addressed technology issues within the framework of indexing and retrieval by the established statistical techniques (though using adjacency phrases as well as single terms), and the fact that TMC's performance was not better is *no* criticism. TMC could expect to benefit, within its technological framework, from the ways of improving performance established by TREC in general. TMC's baseline *P* is in the group ≥ 35 for adhoc and ≥ 40 for routing (for document cutoff). Thus, we see in Table 1 that the best adhoc group, at ≥ 55 , has achieved a performance improvement of 57%, and the best routing group, at ≥ 60 , has increased performance by 50%. These $\ggg$ (or, perhaps $\gggg$) improvements can legitimately be called *striking*. They also represent, from the user's point of view, a real performance difference: getting 16.5 relevant documents as opposed to 10.5 (adhoc) or 18 instead of 11 (routing) would be welcome in many circumstances. On the other hand, it has also to be borne in mind that many teams achieved adhoc 50 or routing 55% *P*, so the difference between their performance and the best, while a noticeable (10%) difference by my earlier definition, would only bring an extra 1-2 relevant documents, given some 15 already.

The top group teams, and those near them in the group below, are primarily from the statistical camp, and although they apply different specific combinations of IR device, have much in common, tending to share a use of compound terms, rich weighting schemes, careful tuning for numerical values, and in Amherst's case in particular, a deliberate attack on the characteristic features of the TREC requests.

5. ASSESSMENT

From one point of view the fact that, here again, relatively straightforward statistical approaches work perfectly reasonably confirms previous results in the field. This is important because the main criticism of previous research has been its comparatively small scale; here in the 'long & large' case, the results still hold. It is, however, still necessary to consider whether there may have been favourable biasing conditions in the TREC case and also, more generally, how the TREC case affects interpretation and application of the test results. This needs to be done in relation both to the TREC data properties and to the evaluation methodology.

5.1 TREC data effects

The TREC data properties of importance here are the heterogeneous document sets, elaborate queries, and form of relevance need, plus the fact that all the data- document sets, queries, and assessments - were 'non-natural'.

The high quality of the requests may have been of considerable significance both in allowing a high level of performance and in permitting it with 'routine' statistically based techniques. For instance, in earlier IR tests, large-scale query expansion was found *damaging*, so although it is not completely clear whether the gains from it in TREC were due to the queries providing better anchoring or to superior expansion criteria, it is likely that the former was important. The high request quality, and especially the carefully developed Concepts fields, may equally mean that sophisticated query derivation methods were not required. The next phase of the TREC programme, in TREC-3, is intended to test alternative, much shorter and 'more ordinary' requests.

Again, it is possible that the large relevance sets, perhaps associated with the character of the TREC relevance needs, may have made it easier to retrieve relevant documents than is usual, and future tests with other relevance criteria and sets are needed.

Thus, for both reasons, although statistically based approaches may be as good as others, TREC may have reached a much higher performance level than could be expected with poorer starting requests (cf. also Blair & Maron, 1985), with knockon effects from relevance supply and query modification.

The non-natural document sets may be unfavourable to TREC, since in ordinary circumstances data sets may be explicitly distinguished and offered for user preview, on global characterisations. The comparative lack of variety in the TREC requests, on the other hand, may have been favourable to TREC in allowing more tuning than would ordinarily be feasible. There is no doubt that the 'ordinary user' has not hitherto been present in TREC, and needs to be in future tests.

5.2 TREC methodology effects

The main issues here are the dual use of requests for routing and adhoc retrieval, the form of the routing tests, and the lack of user participation in reacting to routed documents or in conducting adhoc searches. As mentioned earlier, the lack of 'reality' with respect to users is more serious now than in earlier tests, when batch searching was also normal in operational systems.

The most important consequence of the first point was mentioned in the last section, since it is very unlikely that even if routing requests are honed, adhoc ones will be. Thus, TREC adhoc performance, though tower than routing, may still be far too high. At the same time, routing performance may also be too high because of the large amounts of systematically obtained training data. Although in principle streamed routing would allow the gathering of many assessments with little effort, it is not clear whether the assessments would be consistent over time, or in fact very voluminous.

The inferences to be drawn about real routing system design from the TREC tests are also not clear, since genuine streaming and 'online' allocation of documents to users may create different retrieval conditions calling for, for example, different search strategies or query modification. For instance, it is not obvious what the effects would be for sought recall or precision. This point refers both to the more formal aspects of test design and to the lack of direct user involvement in the TREC case; future TRECs need to address at least the former, and a modest beginning will indeed be made here in TREC-3, through the application of thresholds in evaluation.

However, the most important weakness of TREC so far has been the lack of end-user involvement in searching, since this means that there has been no investigation of the way users respond to the 'long & large' conditions and, in particular, if they do not cope well, what strategies for helping them would be most effective. This applies, for instance, to whether adequate starting requests can be formulated, or whether effective bootstrapping in relevance feedback can be guaranteed. The planned TREC-3 interactive searches are thus extremely important as a check on the results obtained so far, though these tests too will only be a first attack on this issue, since they will not involve 'real' users and will be subject to the artificialities and restrictions that follow from the design need for practically viable and reasonably controlled evaluation.

5.3 Technology issues

In a general way, TREC put a heavy stress on virtually all the participating teams' technology, and it is a striking comment on computational progress, and its implications for automatic indexing and retrieval, that even quite small teams were able to carry out these demanding tests on such large volumes of material.

Computational strategies were generally organised round the need to conduct many experiments, so not very strong conclusions should be drawn about the operational implications of any of the indexing and retrieval strategies studied. However, it is worth noticing, on the one hand, that some approaches required very substantial processing effort (e.g., Bellcore's computation of latent semantic vectors), and on the other, that a few teams deliberately studied the computational efficiency implications of their indexing and searching strategies, with operational performance in mind. These included TMC.

working with a mainstream statistical approach, TRW, and CITRI; ETH and ERIM explored reductive coding schemes.

Looking at the TREC tests from this efficiency point of view, it is fair to say that they show that some quite costly approaches are not wholly out of court, though significant sealing up, or frequent revision, of an indexing vocabulary does present major problems. However, it is also possible to draw the conclusion that, should computational pressures prohibit significant pre-processing of vocabulary or documents, there are viable IR strategies that do not require these.

5.4 Overall research findings

If we now stand back and ask what *general findings* for IR as a whole emerge from the TREC results, subject to the matters just discussed, these seem to be as follows. Note that given the varied masses of test detail, these findings are necessarily broad statements, and ones reflecting majority generalisations rather than universal truths. They are also intended to apply to both adhoc and routing modes, though possibly more strongly to one rather than the other, either for intrinsic reasons or because there is more evidence in one case than the other.

On this basis my findings are:

1. that a reasonable level of performance for large sets of full-text documents can be achieved by statistically motivated techniques;
2. that within this type of approach great improvements over the simple 'baseline' can be achieved;
3. that these improvements are apparently promoted by combinations of data and process that are not necessarily individually very helpful;
4. that these approaches require tailoring to generic collection properties, and typically also tuning to specific collection characteristics;
5. that computationally, system parametrisation is practicable, and actual retrieval can be very efficient;
6. that the quality of requests (and hence queries) appears very important; and
7. that systematically exploited feedback relying on relevance information is valuable.

As an alternative way of obtaining a findings summary, we can make use of the idea of *benchmark* rather than baseline performance. This also has the advantage of allowing for the possibility that the TMC results we have forcibly adopted as a baseline are unusually low for some hidden reason. Thus, if we take the kind of slightly richer statistical approach, using several types of weighting information and extending to simple phrases, that is exemplified by SMART and has established performance credentials, as a generic benchmark, this allows a sharper formulation of the TREC findings. This benchmark appears in the middle block for adhoc in Table 1, and it is evident that it is hard to do conspicuously better in adhoc searching. With routing, on the other hand, it is, not surprisingly, possible to make larger gains using relevance information over the notional baseline or benchmark represented by the adhoc performance for similar collection data.

5.5 Overall methodology lessons

The TREC tests to date are essentially just a rerun, on a much larger scale, of the standard evaluation procedures used for many years. As noted earlier, these involve a high degree of abstraction over system behaviour. However, both the standard *evaluation results* given for the submitted search outputs in the Proceedings, and Harman's own analyses in these, show how easy it is nowadays to manipulate the raw output data from a variety of points of view. Thus, although the official evaluation results have deficiencies, for example, those associated with recall cutoff, and significance testing is needed, these data are still very informative. Moreover, the sheer scale of the tests makes it difficult to carry out the detailed failure analyses that are ideally required, just as it also makes it difficult to conduct meaningful investigations of the effects of end-user needs or behaviour.

However, whereas it may be very hard to conduct adequate diagnostic tests on this scale, in order to achieved properly controlled comparisons, the volume of test and evaluation results does to some extent

justify the rather coarse view implicit in the whole, namely, that it's what you get, not how you get it, that counts.

Thus, perhaps the single most important observation about TREC to date, relevant both to research testing and operational design, is that single 'errors' can have quite striking effects (cf. the report by Bellcore), and that it is necessary to do exhaustive testing (and learning) to establish reliable and consistent performance.

6. THE FUTURE

Here the focus is (a) on TREC as an *evaluation exemplar* and (b) on TREC as a *prototyping guide*. The first refers both to the future of TREC itself, and to its role in IR experiment in general; the second to how both the TREC findings as such and the lessons to be drawn from the TREC enterprise as a whole may be applied to operational systems.

6.1 Prospective and possible TRECs

As already noted, TREC-3 will begin to tackle other request types and interactive searching, as well as more realistic routing. All of these will need further development. It will also be important in the longer run to engage with a wider range of document genres, and with more sensible database partitions, as well as with, on the other hand, more heterogeneous aggregative resources like those flooding the Internet. Looking ahead, the related TIPSTER programme is already working with more than one language, and for future TRECs, both tests of methods on different languages and genuinely multilingual retrieval are needed. There is also a need to do further scaling-up tests.

For IR experiment more generally, TREC has roles both in providing some performance benchmarking and an evaluation reference paradigm and further, amplifying these, can and should play an important part in fostering research when the actual TREC Collection, test results, and evaluation results are made available, in machine-readable form, to the community. Here at last are worthy, if not ideal, successors to the exhausted test collections of former times. In this context, the Questionnaire responses provided by the teams give valuable supporting detail of a kind too often lacking in the past. Though it is not always easy to characterise individual approaches using some standard set of questions, the attempt both to gather explicit statements about detailed choices and processes, and to do this in a way allowing comparisons on specific points, should be a stimulus to better reporting practice and test methodology in the field.

6.2 Applying the TREC findings to systems

The main outcome should be a more widespread and wholehearted implementation of the mainstream statistical approach to operational systems. This must, however, be subject to tailoring and tuning, and to proper support, especially for adhoc requests, for the user in initiating and responding to the search process.

7. CONCLUSION

The major claim made for the ARPA programmes in general is that they hustle research communities very effectively, leading to significant and rapid improvement in task performance. This comes from simple time pressure and from the strongly competitive zeitgeist, but also, especially for a programme with as many teams as TREC, from the interactive and communicative synergy that working together in a common framework promotes. One of the major issues is generalisation, and here some ARPA tasks have broader implications than others. TREC is perhaps more generalisable than some other tasks, in part because it is continuing a long research tradition, which other tasks have lacked, and so can confirm past results as opposed to just producing initial ones.

BEST COPY AVAILABLE

The IR community should indeed be enormously grateful to TREC. It has revitalised IR research and also demonstrated the importance of text retrieval to wider communities (e.g., those engaged with natural language processing). The attention paid to the evaluation in all aspects of its design and execution has set high standards for future work. More specifically, the IR community owes Donna Harman a huge debt for driving the whole.

Acknowledgement - I have benefitted, as a member of the TREC-3 Programme Committee and as an Advisor to the City University team, from my involvement in TREC, and from the many discussions with Donna Harman and others to which this had led. I am also grateful to City University for the travel support that has allowed me to attend TREC meetings. Finally, I should thank my referee for valuable, constructive comments on the draft of this paper.

REFERENCES

- Bates, M. (Ed.) (1993). Human language technology (Proceedings of the ARPA Workshop, 1993). San Mateo, CA: Morgan Kaufmann.
- Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28, 285-299.
- Buckley, C. (1993). The importance of proper weighting methods. In M. Bates (Ed.), *Human language technology*. San Mateo, CA: Morgan Kaufmann.
- Gallant, S.I. (1991). A practical approach for representing context and form performing word sense disambiguation using neural networks. *Neural Computation*, 3, 293-309.
- Galliers, J.R., & Sparck Jones, K. (1993). Evaluating natural language processing systems. Technical Report 291, Computer Laboratory, University of Cambridge. (Gzipped copy is available from the anonymous FTP server <ftp://cl.cam.ac.uk>, as the file `TR291-ksj-jrg-evaluating-nl-systems.ps.gz` in the directory `reports`; the file is compressed so must be transferred in binary.)
- Harman, D.K. (Ed.) (1993). The First Text REtrieval Conference (TREC-1). NIST Special Publication 500-207, National Institute of Standards and Technology. Gaithersburg, MD.
- Harman, D.K. (Ed.) (1994a). The Second Text REtrieval Conference (TREC-2). NIST Special Publication 500-215, National Institute of Standards and Technology. Gaithersburg, MD.
- Harman, D.K. (1994b). Review of TREC, *Information Processing & Management*, 31, 271-190.
- Lewis, D.D., & Sparck Jones, K. (in press). Natural language processing for information retrieval. Technical Report 307, Computer Laboratory, University of Cambridge. *Communications of the ACM*.

APPENDIX A

Key to organisations

ADS	Advanced Decision Systems
Amherst	University of Massachusetts at Amherst
Bellcore	Bellcore
Berkeley	University of California at Berkeley
CITRI	Collaborative LT Research institute, University of Melbourne, Royal Melbourne Institute of Technology
City	City University, London
CMU	Carnegie Mellon University
Conquest	ConQuest Inc
Cornell	Cornell University
CUNY	Queens College, City University of New York
Dalhousie	Dalhousie University
Dortmund	University of Dortmund
ERIM	Environment Research Institute of Michigan
ETH	Swiss Federal Institute of Technology
Florida	University of Central Florida
GE	General Electric R & D Centre
HNC	HNC Inc
IDS	Institute for Decision Systems Research
Illinois	University of Illinois at Chicago
Mead	Mead Data Central
NYU	Courant Institute, New York University
PRC	PRC Inc
Rutgers	Rutgers University
SEC	Systems Environment Corporation
Siemens	Siemens Corporate Research Inc
Syracuse	University of Syracuse
TMC	Thinking Machines Corporation
TRW	TRW Systems Development Division, Paracel Inc
UCLA	University of California at Los Angeles
Verity	Verity Inc
VT	Virginia Tech

This paper first appeared in *Information Processing and Management* Vol. 31, No. 3, 1995 p. 291-314 and has been reproduced here with the kind permission of the original publishers, Pergamon-Elsevier.

QUESTIONS AND ANSWERS

"Be practical" exhorts Carl Keren in his letter to JASIS and is taken to task by the recently late Gerard Salton. A Conference on new methods in information retrieval organised early in 1996 by Infonortics Ltd. was introduced by Ev Brenner. He dedicated the Conference to the memory of Salton with the observation that most of the products and approaches exhibited, discussed and demonstrated at the Conference showed that his decades of research had culminated in a wide range of his ideas firmly coming to the market place. So Carl Keren would be encouraged, but perhaps not fully satisfied. The paper by Weiner and Liddy may give us a glimpse of the promised land but with that degree of computer power not yet widely accessible. Answers, more questions, more answers and tests - the cycle continues.

Alan Gilchrist

BEST COPY AVAILABLE

137

ON INFORMATION SCIENCE

Carl Keren

National Center of Scientific and Technological Information (COSTI)
P.O. Box 20125, Tel-Aviv 61 200, Israel

Sir:

JASIS, Vol. 34, No. 1, has only now reached me and following your invitation I would like to comment on the editorial "Information Science, As a Name and a Subject"¹.

I strongly endorse the opinion that most readers of *JASIS* (and of similar journals) are practitioners and I believe it to be true that most writers are researchers. The question, however, arises - what are the writers (the researchers) writing about. Do they write about research which advances the state-of-the-art and which, eventually, contributes to a better understanding of natural laws or the behavioral attitudes of society? Do they write about advances which will, eventually, enable the practitioner to develop a better system, or machine or even a placebo, anything which would create a new base line (however, insignificant) from which the art can further evolve?

In the natural sciences, and in many of the behavioral sciences, researchers make studies of past achievements, formulate a hypothesis, go on to develop a theory and then try to validate their concepts through experimentation and firm data. At the end of this chain new knowledge emerges and existing knowledge is updated or subjected to additional questioning. Nearly constantly, the state-of-the-art of a particular science is pushed a tiny little bit further, a new base line for further progress is opened up for inspection by professional peers and/or for utilization by practitioners. Biologists who would repeatedly rediscover with Pasteur the cause of rabies and the life cycle of its microbes or economists who would endlessly harp back to the Luddites, are not very likely to find a respectable place in their professional journals.

Although research in information science is not, perhaps, that stagnant it would be worthwhile to find out how much of it has really contributed to our body of knowledge and to the methods used by practitioners. I dare say that we will probably be rather disappointed. I did not research this, nor do I have firm data, but on the strength of my own experience, as a practitioner, I have the impression that only very rarely useful and nontrivial information emanates from the research published in our professional journals, including those who strive to set high standards.

Some of the research which is reported is intellectually entertaining, much more has a "déjà vu" sheen. Some reports are festooned with mathematics (my late friend Y. Bar-Hillel had always cautioned that an information scientist who needs mathematics to describe his concepts quite probably knows little of either information science or mathematics), others are obvious manifestations of degree mills or of the publish-or-perish syndrome. A great many researchers seem to be ignorant of previously published work, especially if it has been done in another country. (Just imagine a chemist who would do so, he would soon be shown the errors of his ways). Much so-called "research" milks the results of some questionnaire or casual survey ad infinitum and frequently ad nauseam. The emergence of a new concept is a very rare experience.

Where does all this leave information science? Is it a name with some recognizable body attached to it or is it a subject which is a figment of someone's imagination? Is the body real, has it a substance which we are learning to know and to recognize, in whatever guise it appears, do we know how that body

grows and develops and which laws govern its behavior? Is it a subject whose contents we can define and whose borders are known at any specific time. I believe that the reply to all of these questions is as vague today as it was years ago. To reply properly to Dr. Simkin's very valid questions some good answers had better be forthcoming soon. That we do not have them yet is surely, at least partly, the fault of the research establishment and its lack of communication with and feedback from the technologists. This lack of communication is probably also the fault of technologists, many of whom do not have the training, the motivation and the scientific inclination to demand more than they are offered.

Journal editors could help by applying much more stringent refereeing processes which would enable us to select from the quantities we are offered, the qualities which we cannot afford to do without. Such qualities should progressively build up the substance of information science and thereby create a subject worthy its name. At best, we would start to understand the knowledge transfer processes and be able to do something with and about them, at worst we will get to know that we are chasing a phantom, a name which has no real substance or subject contents, at least in the sense we use it.

Carl Keren
National Center of Scientific and Technological Information (COSTI)
P.O. Box 20125,
Tel-Aviv 61 200, Israel

1. Meadow, C. T. "On Information Science, As a Name and a Subject." *Journal of the American Society for Information Science*. 34(1):3-4; 1983.

KEREN: This paper was first published in *Journal of the American Society for Information Science: JASIS* 35(2) 137, 1984 (ISSN 0002-8231) and has been reproduced here with the kind permission of John Wiley & Son, Inc.

BRIEF COMMUNICATION

A NOTE ABOUT INFORMATION SCIENCE RESEARCH¹

Gerard Salton

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

This note deals with the relationship between information science research and practice. The impression that the field is moribund and that the research output is uniformly inferior is not supported by an examination of the information retrieval literature.

A provocative note by Carl Keren appeared in a recent issue of *JASIS* under the title "On Information Science"¹. The note raises interesting questions about the relationship between information science research and practice. Keren asks four questions (slightly paraphrased):

- (a) Do researchers in information science write about research that advances the state of the art?
- (b) How much of the research output has really contributed to our body of knowledge?
- (c) Is information science a name with a recognizable body attached to it? Is it a subject whose contents we can define?
- (d) Is there a lack of feedback between researchers and technologists?

Questions such as these are of course not new, and they are reflective of legitimate concerns. In fact, I share some of Mr. Keren's misgivings about the quality and quantity of acceptable research work in our field. But we part company when Keren, by implication, answers all his questions in the negative, and specifically blames the researchers for what he considers the unsatisfactory state of information science work. In stating his arguments, Keren makes valid points about research based on questionable survey output, about lack of knowledge of the literature, about rediscovery of the obvious, and the like. Unfortunately, the note is also replete with remarks of doubtful relevance and with unsubstantiated claims about the inferior state of information science compared with that of other natural sciences. Taken as a whole, the note presents an antiresearch, if not anti-intellectual image, and requires, I believe, some amendments.

Among the less edifying parts of Keren's note is a completely gratuitous quotation of a remark, purported to have been made by Y. Bar-Hillel, to the effect that information scientists who need mathematics to describe their concepts quite probably know little of either information science or mathematics. To which one replies, first, that the remark has little to do with Keren's concern about the quality or quantity of research in information science, and second that the comment is obviously foolish and should not, in fairness to Bar-Hillel, have been repeated. (Bar-Hillel was an astute observer of our field; but he was partial to sweeping, and sometimes indefensible generalizations which might better be forgotten.) The truth of the matter is that a mathematical approach cannot transform poor information science into good research, and that few people are able to make convincing use of mathematics in the information field. But this does not, of course, imply that all mathematics is extraneous to information science, or that all information scientists who use mathematics are inferior.

¹ Accepted October 16, 1984

What about Mr. Keren's main question? I will beg leave from answering the question about the scope and definition of information science. I just do not find that discussions relating to the definition of disciplines advance the state of the art, and personally I do not have much to contribute. I will then confine myself to the other questions concerning the quality and quantity of research in information science, and the contributions of this research to our body of knowledge and to information science practice. In discussing these questions, Mr. Keren mentions that "I did not research this, nor do I have firm data, but on the strength of my own experience, as a practitioner, I have the impression that only very rarely (does) useful and nontrivial information emanate from the published research...."

Well, I do not have any firm data either, but my own assessment generally agrees with Keren's: Most of the published research in our field is probably not worth doing and ought to be forgotten. I differ with Keren in that I do not jump to the implied conclusion that therefore our researchers are all inferior and that information science research is useless. Nor do I have any evidence that relative to the number of researchers in information science, the proportion of first-rate work in our field is smaller than it is in other fields. The latter point is made explicitly at some length in Keren's note.

As I said, I cannot give conclusive evidence either way, but I can cite the literature. What follows is a list of active research topics pursued by people who generally have something to say. All of these topics belong to that part of information science which may loosely be entitled "text analysis, storage, and retrieval," and the literature citations cover the three most recent years (1981-1983). I emphasize that the list does not constitute a ranking of the best people, or the best topics, or the best papers. It is simply a sampling of recent work that offers interesting perspectives to researchers and practitioners alike. For a different view of the information science field in areas not covered here, the reader may wish to turn to other recent surveys²³.

(1) Under the **vector processing** retrieval strategy, both queries and stored documents are represented by sets of terms without Boolean operators. The vector processing model has been criticized because the implicit assumption is that the terms are independent of each other. Bookstein⁴ has shown, however, that under various stated assumptions the preferred probabilistic retrieval approach and the vector processing systems are closely related. The vector processing model, whose effectiveness is hard to surpass in practice, has also been studied more recently by Wong and Raghavan⁵.

(2) **Probabilistic retrieval** models have been preferred by researchers because they have a proper theoretical base and allow, at least in theory, the use of arbitrary combinations of dependent terms. Two different, well-known, probabilistic models have recently been compared, and a unified model including elements of both models has been proposed by Robertson, Maron, and Cooper^{6,7}. An elegant, theoretically optimal probabilistic model based on the maximum entropy principle has been introduced by Cooper⁸. Croft has described and evaluated procedures for extending the basic probabilistic model by allowing the use of weighted terms for document representation^{9,10}. Yu and co-workers¹¹ have introduced a generalized probabilistic model in which dependent term pairs, term triples, and higher-order term combinations can be included in decreasing order of the importance of the various term combinations. Finally, Wu and Salton¹² have evaluated some practical methods for actually obtaining the needed probabilistic parameters for the case where term independence is assumed.

(3) The probabilistic retrieval approach is theoretically attractive, but until viable methods appear for actually obtaining the probabilistic parameters for dependent-term combinations, the simpler vector processing approaches will be used in practice, or the established **inverted file** procedures based on term list manipulations. In this connection substantial attention has been paid recently to the acceleration and optimization of the similarity computations between queries and documents using inverted file strategies. Smeaton and van Rijsbergen¹⁴ and Murtagh¹⁵ have worked on the optimization of near-neighbor computations usable for query-document comparisons. Noreault and Chatham¹⁶ have applied similar methods to the computation of term-term similarities, and Willett^{17,18} has used optimal inverted file strategies to compute the document-document similarities needed for document clustering. All of these optimized inverted file procedures produce substantial improvements over conventional inverted file methodologies and are directly usable in current operational retrieval environments.

(4) The well-known query reformulation process based on **relevance feedback** uses relevance assessments obtained from retrieval system users in response to earlier information searches. Attar and Frankel¹⁹ have continued working on local (and hence inexpensive) clustering and relevance feedback procedures. Term relationships obtained from a maximum spanning tree of term-term similarities are used in a relevance feedback environment by van Rijsbergen, Harper, and Porter²⁰. All of these automatic query expansion and reformulation methods appear to be eminently suitable for practical implementation²¹.

(5) Operational retrieval systems are currently based on **Boolean query formulations**. The conventional Boolean methodology is rudimentary: no term weighting, no output ranking of retrieved items, no automatic query formulation. A great deal of attention has been devoted by researchers to the refinement of Boolean query manipulations. Bookstein²² has compared different term weighting systems in Boolean query environments. Buell and Kraft^{23,25} and Radecki²⁶ have studied the **fuzzy set** retrieval models which permit the use of weighted document identifiers with Boolean queries. A different extended Boolean system was introduced by Salton, Fox, and Wu²⁷ which is more general than the fuzzy set system and appears to produce large improvements in retrieval output. This system can be superimposed on the currently operating retrieval methods. Regrettably, some mathematical considerations are needed to understand the operations of the extended system.

(6) Fancy **front-end procedures** have been studied by many people and beginnings have been made toward the design of **expert systems** that facilitate the access to the stored collections²⁸⁻³⁴. Marcus²⁹, has designed a flexible system that uses a common query language to access a variety of different bibliographic retrieval systems. The interactive system by Doczkocs³⁰ is based on flexible term weighting and automatic thesaurus display operations. Frei and Jauslin³³ use graphic terminals and multiple window displays to guide the user through the query formulation process using a hierarchical thesaurus display. Finally, Guida and Tasso³⁴ propose an expert-type interactive system in which user responses to appropriately phrased questions are utilized as a guide to the search process.

(7) Bibliographic **citation networks** have been used for many years to characterize various disciplines and literature types. Kwok³⁵ suggests that terms from cited titles be used for document identification purposes, and White and Griffith^{36,37} continue the extensive experimentation with citation network manipulations.

(8) The area of **text passage retrieval**, as opposed to full document retrieval, has not yet come into its own, but it offers interesting prospects as a compromise between document retrieval and question-answering. O'Connor³⁹ has done pioneering work in this area, and Bernstein and Williamson⁴⁰ have built an interesting system for retrieval of text passages from a specially constructed handbook on hepatitis diseases.

(9) The related area of automatic indexing using **linguistic analysis** is being seriously revived by various researchers. For example, Dillon and co-worker^{41,42} use phrase assignment techniques in various text processing experiments, and Sparck Jones and Tait⁴³ describe a sophisticated system for the automatic generation of syntactically (and semantically) acceptable term phrases.

(10) The interesting area of fast text **character matching** - for example, between query and document texts - is treated by Haskin and Hollaar⁴⁴ who propose the construction of special hardware devices for rapid character string processing. Sophisticated **simulation methods** for the construction of representative bibliographic databases are described by Tague and Nelson⁴⁵. Simulated collections simplify the operations of controlled retrieval evaluation experiments. Finally, a large group of people including, for example, Macleod and Crawford⁴⁶, are interested in the design of **mixed search and retrieval** systems that process formatted numeric databases as well as text data in an integrated operation.

All of the foregoing research areas are directly relevant to information retrieval practice in that they point the way to the construction of flexible, interactive retrieval systems which should substantially simplify the collection search and retrieval processes for future generations of information seekers. Many other information processing topics might have been given prominence: for example, recent attempts at automatic abstracting⁴⁸, word fragment encoding experiments, the design of distributed retrieval systems⁵⁰, and the area of systems evaluation. For present purposes, the foregoing list suffices to

demonstrate that an impressive variety of research topics is being pursued by people who have something to offer.

In the face of the relative wealth of interesting work actually available, Mr. Keren's rhetorical questions about "What have you researchers done lately for us practitioners?" seems to miss the point. There are never any shortcuts in bridging the gap between research and application. That is true in our area as much as in all other intellectual areas of endeavor. It is necessary to study the literature; it is necessary to have sufficient know-how to discriminate, and to put matters in context. Eventually the pieces will fit together, and the observer can judge the specifics instead of being forced to rely on superficial impressions and generalizations. In our field, as in every other, it is necessary to know the field before being able to contribute.

I see no reason, however, to resign. There has been substantial progress in information system design these past 20 years. No doubt, the changes will accelerate in the years to come. The field will evolve even more rapidly if the practitioners were to stop blaming the research side and asked instead "What have we practitioners lately done for ourselves?"

REFERENCES

1. Keren, C. "On Information Science." *Journal of the American Society for Information Science*. 35(2):137; March 1984.
2. Kochen, M. "Information Science Research: The Search for the Nature of Information." *Journal of the American Society for Information Science*. 35(3):194-149; May 1984.
3. Herner, S. "Brief History of Information Science." *Journal of the American Society for Information Science*. 35(3):157-163; May 1983.
4. Bookstein, A. "Explanation and Generalization of Vector Models in Information Retrieval." In: Salton, G.; Schneider, H.J., eds. *Research and Development in Information Retrieval, Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:118-132.
5. Wong, S.K.M.; Raghaven, V.V. "Vector Space Model of Information Retrieval - A Reevaluation." In: van Rijsbergen, C.J., ed. *Research and Development in Information Retrieval*. Cambridge, England: Cambridge University Press; 1984:167-186.
6. Robertson S.E.; Maron, M.E.; Cooper, W.S. "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval." *Information Technology: Research and Development*. 1(1):22; January 1982.
7. Maron, M.E. "Associative Search Techniques Versus Probabilistic Retrieval Models." *Journal of the American Society for information Science*. 33(5):308-310; September 1982.
8. Cooper W.S.; Huizinga, P. "The Maximum Entropy Principle and its Application to the Design of Probabilistic Retrieval Systems." *Information Technology: Research and Development*. 1(2):99-112; April 1982.
9. Cooper, W.S. "Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness." *Journal of the American Society for Information Science*. 34(1):31-39; January 1983.
10. Croft, W.B. "Document Representation in Probabilistic Models of Information Retrieval." *Journal of the American Society for Information Science*. 32(6):451-457; November 1981.
11. Croft, W.B. "Experiments with Representation in a Document Retrieval System." *Information Technology: Research and Development*. 2(1):1-22; January 1983.
12. Yu, C.T.; Buckley, D.; Lam, K.; Salton, G. "A Generalized Term Dependence Model in Information Retrieval." *Information Technology: Research and Development*. 2(4):129-154; October 1983.
13. Wu, H.; Salton, G. "The Estimation of Term Relevance Weights Using Relevance Feedback." *Journal of Documentation*. 37(4):194-214; December 1981.
14. Smeaton, A.F.; van Rijsbergen, C.J. "The Nearest Neighbor Problem in Information Retrieval." *Proceedings of Fourth International Conference on Information Storage and Retrieval "SIGIR Forum"*. 16(1):83-87; Summer 1981.
15. Murtagh, F. "A Very Fast Exact Nearest Neighbor Algorithm for Use in Information Retrieval." *Information Technology: Research and Development*. 1(4):275-283; October 1982.

16. Noreault, T.; Chatham, R. "A Procedure for the Estimation of Term Similarity Coefficients." *Information Technology: Research and Development*. 1(3):189-196; July 1982.
17. Willett, P. "Document Clustering Using an Inverted File Approach." *Journal of Information Science*. 2:223-231; 1980.
18. Willett, P. "A Fast Procedure for Calculation of Similarity Coefficients in Automatic Classification." *Information Processing and Management*. 17(2):53-60; 1981.
19. Attar, R.; Frankel, A.S. "Experiments in Local Metrical Feedback in Full-Text Retrieval Systems." *Information Processing and Management*. 17(3):115-126; 1981.
20. van Rijsbergen, C.J.; Harper, D.J.; Porter, M.F. "The Selection of Good Search Terms." *Information Processing and Management*. 17(2):77-91; 1981.
21. Pietilainen, P. "Local Feedback and Intelligent Automatic Query Expansion." *Information Processing and Management*. 19(1):51-58; 1983.
22. Bookstein, A. "A Comparison of Two Systems of Weighted Boolean Retrieval." *Journal of the American Society for Information Science*. 32(4):275-279; July 1981.
23. Buell, D.A.; Kraft, D.H. "Threshold Values and Boolean Retrieval Values." *Information Processing and Management*. 17(3):127-136; 1981.
24. Buell, D.A.; Kraft, D.H. "A Model for a Weighted Boolean Retrieval System." *Journal of the American Society for Information Science*. 32(3):211-216; May 1981.
25. Buell, D.A. "A General Model of Query Processing in Information Retrieval Systems." *Information Processing and Management*. 17(5):249-262; 1981.
26. Radecki, T. "Reducing the Perils of Merging Boolean and Weighted Retrieval Systems." *Journal of Documentation*. 38(3):207-211; September 1982.
27. Salton, G.; Fox, E.A.; Wu, H. "Extended Boolean Information Retrieval." Technical Report 82-511, Department of Computer Science, Cornell University, August 1982; also in *Communications of the ACM*. 26(11):1022-1033; November 1983.
28. Meadow, C.T.; Hewett, T.T.; Aversa, E.S. "A Computer Intermediary for Interactive Data Base Searching." *Journal of the American Society for Information Science*. 33(5):325-332; September 1982; also 33(6):357-364; November 1982.
29. Marcus, R.S. "Experimental Comparison of the Effectiveness of Computers and Human as Search Intermediaries." *Journal of the American Society for Information Science*. 34(6):381-404; 1983.
30. Doczkocs, T.E. "From Research to Application: The Cite Natural Language Information Retrieval System." In: Salton, G.; Schneider, H.J. eds. *Research and Development in Information Science, Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:251-262.
31. Heine, M.H. "A Simple Intelligent Front End for Information Retrieval Systems Using Boolean Logic." *Information Technology: Research and Development*. 1(4):247-260; October 1982.
32. Pollitt, A.S. "End User Touch Searching for Cancer Therapy: A Rule Based Approach." *Proceedings of Sixth Annual Conference on Research and Development in Information Retrieval, SIGIR Forum*. 17(4):136-145; Summer 1983.
33. Frei, H.P.; Jauslin, J.F. "Graphical Presentation of Information and Services: A User-Oriented Interface." *Information Technology: Research and Development*. 2(1):23-42; 1983.
34. Guida, G.; Tasso, C. "IR-NLI: An Expert Natural Language Interface System." Conference on Applied Natural Language Processing, Association for Computational Linguistics, Santa Monica, CA, 1983. PP. 31-38.
35. Kwok, K.C. "A Document-Document Similarity Measure Based on Cited Titles and Probability Theory and its Application to Relevance Feedback Retrieval." In: van Rijsbergen, C.J. ed. *Research and Development in Information Retrieval*. Cambridge, England: Cambridge University Press; 1982.
36. White, H.D.; Griffith, B.C. "Authors as Markers of Intellectual Space: Cocitation in Studies of Science, Technology and Society." *Journal of Documentation*. 38(4):255-272; December 1982.
37. White, H.D.; Griffith, B.C. "Author Cocitation: A Literature Measure of Intellectual Structure." *Journal of the American Society for Information Science*. 32(3):163-171; May 1982.
38. Norma, E. "Untangling Citation Networks." *Information Processing and Management*. 18(2):43-53; 1982.
39. O'Connor, J. "Citing Statements: Computer Recognition and Use to Improve Retrieval." *Information Processing and Management*. 18(3):125-131; 1982.
40. Bernstein, L.M.; Williamson, R.E. "Testing of a Natural Language Retrieval System for a Full

- Text Knowledge Base." *Journal of the American Society for Information Science*. 35(4):235-247; July 1984.
41. Dillon, M.; Gray, A.S. "Fasit: A Fully Automatic Syntactically Based Indexing System." *Journal of the American Society for Information Science*. 34(2):99-108; 1983.
 42. Dillon, M. "Thesaurus-Based Automatic Book Indexing." *Information Processing and Management*. 18(4):167-178; 1982.
 43. Sparck Jones, K.; Tait, J.I. "Automatic Search Term Variant Generation." *Journal of Documentation*. 40(1):50-66; March 1984.
 44. Haskin, R.L.; Hollaar, L.A. "Operational Characteristics of a Hardware-Based Pattern Matcher." *ACM Transactions on Database Systems*. 8(1):15-40; March 1983.
 45. Tague, J.; Nelson, M. "Simulation of Bibliographic Databases Using Hyperterms." In: Salton, G.; Schneider, H.J., eds. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:194-208.
 46. Macleod, I.A.; Crawford, R.G. "Document Retrieval as a Database Application." *Information Technology: Research and Development*. 2:43-60; 1983.
 47. Schek, H.J. Methods for the Administration of Textual Data in Database Systems," in: Oddy, R.N.; Robertson, R.E.; van Rijsbergen, C.J.; Williams, P.W., eds. *Information Retrieval Research*. London: Butterworths; 1981:218-235.
 48. Paice, C.D. "The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-Indicating Phrases." In: *Information Retrieval Research*. London: Butterworth; 1982:172-191.
 49. Kropp, K.; Walch, G. "A Graph Structured Text Field Index Based on Word Fragments." *Information Processing and Management*. 17(6):363-376; 1981.
 50. Moulinoux, C.; Faure, J.C.; Litwin, W. "Messidor: A Distributed Information Retrieval System." In: Salton, G.; Schneider, H.J., eds. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 1983:146:51-61.
 51. Bollmann, P. "Two Axioms for Evaluation Measures in Information Retrieval." In: van Rijsbergen, C.J., ed. *Research and Development in Information Retrieval*. Cambridge, England: Cambridge University Press; 1984:233-246.

SALTON: This paper was first published in *Journal of the American Society for Information Science*: JASIS 36(4) 268-271, 1985 (ISSN 0002-8231) and has been reproduced here with the kind permission of John Wiley & Son, Inc.

UNANSWERED QUESTIONS IN THE DESIGN OF CONTROLLED VOCABULARIES

Elaine Svenonius

Graduate School of Library and Information Science, University of California at Los Angeles,
Los Angeles, CA 90024, USA

The issue of free-text versus controlled vocabulary is examined in this article. The history of the issue, which is seen as beginning with the debate over title term indexing in the last century, is reviewed and then attention is turned to questions which have not been satisfactorily addressed by previous research. The point is made that these questions need to be answered if we are to design retrieval tools, such as thesauri, upon a national basis.

INTRODUCTION

The question of the effectiveness of controlled vocabularies in information retrieval has been raised repeatedly in the literature of library/information science. That this is so suggests that its answer is neither obvious nor trivial, but depends on a number of variables. Foremost among these are what might be called intrinsic variables:

- The nature of the controlled vocabulary, e.g., its size and the levels of control recognized.
- the nature of the subject discipline, particularly as to its terminology: how ambiguous it is, how much terminological consistency it exhibits, and how predictable its naming of concepts is.

There are then external or situational variables. First:

- the nature of the retrieval system, particularly whether it is manual or online and, among the latter, the options possible in designing search strategies.

A good tool can be used clumsily. On the other hand, a task may be accomplished easily without using the tool designed for it. Here external variables of a behavioral kind come into play:

- the skill of the indexer who selects terms from controlled vocabularies to describe documents.
- the skill of the searcher who selects terms to develop search strategies.
- users' retrieval requirements, particularly as these are defined in terms of precision and recall.

In making a decision of whether or not to construct a controlled vocabulary, questions of cost must be considered. Thus, variables not affecting, but associated with, retrieval performance are

- the costs of constructing controlled vocabularies and the costs of searching databases.

An experiment sophisticated and large enough to control all of the above variables has never been conducted and probably never will. Indeed, it is doubtful whether certain of the key concepts that would figure in such an experiment could be operationally defined, for instance, the concept of relevance. But this does not mean research should not be attempted; quite the contrary, to design controlled vocabularies we need to know more about their retrieval effectiveness. The purpose of this paper is to review what we know about the retrieval effectiveness of controlled vocabularies and in so doing raise questions that suggest areas for research.

The purpose of a controlled vocabulary is to redress certain retrieval problems caused by the use of natural language in retrieval. Simply stated, these are the problems of homonymy and synonymy. The first occurs when a given word or phrase has different referents or meanings. Thus, a person searching on the term Drums, meaning musical instruments, might well retrieve material on ear drums or on the fish called drums. The retrieval situation is characterized as exhibiting poor precision.

The presence of synonymy in natural language can also affect retrieval adversely. Synonyms occur when two words share the same meaning, an obvious instance of which is when a given referent is represented by different names: Bulbous domes and Onion domes. The user searching under one of the two terms might not find material indicated by the other, a retrieval situation characterized as exhibiting poor recall.

A controlled vocabulary is designed to enhance the precision and recall of a search by straightening out the crisscross of many-one and one-many relationships between words and their referents. A controlled vocabulary, then, can be regarded as controlled to the degree that it incorporates devices to accomplish this task. There are various kinds of control, e.g., homonym and synonym control, and, with respect to the latter, various degrees of control. Synonymy, strictly interpreted, represents only one step in a process of normalization. Thus, a first step in the normalization process is to regard as equivalent orthographic variations of a given word: Database, Data-base and Data base, or Experiment and Experiments. This type of normalization is usually accomplished by stating some simple rules about how hyphens are to be regarded and how singular and plural forms of a term are to be treated. A second step in the normalization process is to regard as equivalent lexical variants, related word forms, such as Analyzing, Analysis, and Analytic. A third and much larger step is to regard as equivalent terms that are synonymous in meaning. These three steps in imposing control make use of equivalence relationships. However, a given concept can be represented at different levels of generality or from different points of view and, in so doing, can cause retrieval failures. Therefore, most thesauri, and all classifications, go beyond strict equivalence-type relationships and impose successively greater control by incorporating (step 4) hierarchical relationships and (step 5) related terms relationships.

Controlled vocabulary terms are variously called *index terms*, *descriptors*, *subject headings* and, somewhat erroneously, *keywords*. Terms not belonging to a controlled vocabulary are called *free-text terms*, *natural language terms* and, again, *keywords*. *Free-text* is also used as an adjective to describe a type of searching, viz., searching that can be performed without the constraint of having to translate one's own vocabulary into the vocabulary used by a particular system. Most free-text searching, however, involves a modicum of vocabulary control, imposed by the users of the system; for instance, users in their search strategies employ various forms of word stemming; they form equivalence classes of synonymous terms using the Boolean "OR" device, and, on a more advanced level, they make use of related terms that can be generated automatically.

HISTORY

The issue of free-text versus controlled vocabulary has a long history, which can be divided into three eras. Era One began in the last century with the increasing popularity of title-term indexing. Title-term or title-catchword indexing was a precursor of the modern keyword indexing and its search correlate, free-text searching. In the nineteenth century the dominant controlled vocabularies were classification schemes; these were used to order bibliographic records in classified or classed catalogs. Classed catalogs began to prove troublesome as library collections grew. Users complained that although they were able to find a group of books published within a scientific discipline, it was difficult to find a given title.¹ In modern parlance, precision was bad. They also complained that the classed catalog used a vocabulary familiar only to the highly educated. How would the man on the street wanting information about the badger be expected to know that he should look under Science: Division--Natural History: Subdivision - Zoology: Group - Vertebrates: Class -Mammals: Subclass - Monodelphia: Section - Carnivora, and so on.²

The proposed solution to the problems of precision and common usage endemic to classed catalogs was the use of an uncontrolled vocabulary derived from the titles of books. While using title catchwords as points of entry for an index or catalog is a practice that stretches far into the past, favoring this approach for the express purpose of improving precision, beyond what could be achieved by classed catalogs, was not articulated until about the middle of the nineteenth century. A leading proponent of title-term indexing was Samson Low, who in 1854 produced an index to the titles in the *British Catalogue of Books*. He wrote: "It is hoped by following out the author's own definition of his books, and presenting a CONCORDANCE OF TITLES [to] combine both of these advantages (i.e., to find a given title as well as a group of books published within a scientific discipline.)"³ . Perhaps the name most often associated with title term indexing is that of Low's assistant, Crestadoro, who in his *Catalogue of Books in the Manchester Free Library* entered books under each of the significant terms in their titles⁴.

Then, as now, controversy arose over the effectiveness of uncontrolled vocabularies. Charles Ammi Cutter, in marshalling his arguments for a system of alphabetic subject headings, is credited with having dealt the death blow to title-term indexing⁵. Cutter objected to title terms on two grounds: (1) titles may be unintelligible or striking or fanciful and, thus, not express the true subjects of the works they name, and (2) works on precisely the same subject would be separated if the phraseology of their titles were different: *Insects* and *Entomology*; or *Free Trade*, *Protection*, and *Tariff*⁶. Actually, Crestadoro was aware of these objections and to meet them proposed enhancing titles that were not informative and making cross references to link synonyms. But he did not fully implement these proposals or follow out their implications. This was left to Cutter. With the introduction of the alphabetic subject approach, of which the prototype was the *Library of Congress Subject Headings*, the controversy over controlled versus uncontrolled vocabularies subsided for over half a century: it was taken for granted that subject access should be through some form of controlled vocabulary.

Era Two in the free-text versus controlled vocabulary controversy arrived with the advent of the computer and the promise it offered of a quick and effective form of derived indexing. In 1959, Hans Peter Luhn introduced keyword in context indexing (KWIC), a mechanized form of derived indexing wherein the potential ambiguity of the uncontrolled terms was resolved by displaying them in context. Because it does not make use of vocabulary-control devices, other than the supplying of contexts, KWIC is sometimes called "quick and dirty" indexing. Luhn, however, was aware of the limitations of KWIC and suggested it not be used for serious information retrieval, but only for dissemination indexes having the limited purpose of alerting researchers to current information as quickly as possible⁷.

About the same time as KWIC indexes appeared, experiments in Cranfield, England were being directed toward evaluating the performance of various index languages in retrieval⁸. These experiments were notable for introducing operational definitions of retrieval performance in terms of precision (the proportion of retrieved documents that are relevant) and recall (the proportion of relevant documents retrieved). The objective of the second of the Cranfield experiments was to test the effectiveness of vocabulary control. This experiment compared 33 index languages incorporating various degrees of control; its results provided grist to the mill of those doubting whether controlled vocabularies were effective. Most cited was the result that a minimally controlled index language, one in which only synonyms and word endings were normalized, performed as well or better in retrieval than any index language with full vocabulary control. In quick succession, a series of similar experiments followed, reapplying the methodology used in the Cranfield experiments and producing comparable findings⁹.

The retrieval experiments of the 1960s and early 1970s were subject to a great many criticisms¹⁰. One of the most important was that the experimental situations did not accurately reflect reality: The assumed analogy between the laboratory or experimental situation and the real world was particularly suspect in the sets of questions used to represent users' information-seeking behaviour and in the search strategies used to retrieve documents¹⁰. Other criticisms concerned the statistical design of the experiments: Samples were limited to narrow subject areas and the inferential statistics used often were inappropriate (sometimes they were even lacking). There were criticisms that were definitional in nature; these related mostly to the dependent and independent variables. It was commonplace to observe that relevance judgments were subjective. Because precision and recall, the dependent variables in the experiments, were based on relevance assessments, they were subject to the twin

charges of invalidity and unreliability. The independent variable index language also was problematic in that it suffered from definitional ambiguity¹¹. Because of flawed methodology, it was difficult in the retrieval experiments of the 1960s and 1970s to separate out the relative effects of different factors on retrieval performance.

Many accepted the findings of Cranfield II and similar experiments, despite their flawed methodologies. Undoubtedly it was expedient to do so, since the construction of controlled vocabularies entails considerable expense¹². In any case, it could hardly be of concern to database producers if the indexers' burden of controlling vocabulary were transferred to end users, who at the time of search would pay for their possibly amateur thinking at steep online searching costs¹³. (Some search services charge as much as \$3.00 for each search statement entered, i.e., \$3.00 is the cost of forgetting a synonym, overlooking a parenthesis or misspelling a word.)

Era Two, the era of large retrieval experiments, gave way in the middle 1970s to Era Three, in which a different approach was taken to examining the issue of freetext versus controlled vocabularies; not surprisingly, perhaps, different conclusions emerged. In 1976 Barbara Charton wrote a one-and-a-half-page paper in the *Journal of Chemical Information and Computer Sciences*¹⁴, which opened with the question "is a controlled vocabulary necessary?" Charton searched *Chemical Abstracts* (CA), "by hand" and "in likely places" to find papers on correlation analysis. Then she searched for material on correlation analysis using the CA keyword index, an uncontrolled vocabulary constructed of entries selected from the title or text of the documents. Fifty percent of the articles she found by hand she could not find using the keyword index. She wrote a letter complaining about this to the editor of CA. He responded first by observing that the set of keywords used to search the keyword index were not complete for the subject in question, and second by noting, as had Luhn before him, that a keyword index was intended only for current and quick access and that the controlled vocabulary was the proper tool for retrospective searching. He referred to a study he and his colleagues performed which produced the finding that "searching controlled- and uncontrolled-vocabulary files ... gives complementary but not necessarily identical results." He then speculated that a free-text search might be as effective as a controlled-vocabulary search if it were performed by an expert fully conversant with the field being searched; it might even be more effective because "a controlled vocabulary sometimes sacrifices precision in favor of predictability"¹⁵.

The following year (1977) Carrow and Nugent conducted a comparative evaluation of index term versus free-text search methods using the National Criminal Justice Reference database¹⁶. The text searched included document titles, annotations, and abstracts. Examination of 23 search outputs showed that the two search methods had about the same precision performance, but index-term searches produced significantly better recall. The authors observed that the two methods were complementary and hypothesized that the best performance would be achieved by a combination of the two methods.

The idea that free-text searches and searches on controlled vocabulary terms complement one another, and that the best performance could be achieved by the two in combination, was reinforced in the following year (1978) by Henzler in Germany¹⁷. Using a method similar to Charton's, he compared samples of index terms and free-text terms in the CANCERNET database. He found, using two samples of 100 titles, that 35% of all title words had no appropriate meaning equivalents among the descriptors in the controlled vocabulary. Conversely, for 50% of the descriptors assigned to documents no acceptable free-text representations could be found. Henzler concluded: "Thus, the alternative 'free-text or controlled vocabulary' is no longer an alternative: there should always be both free-text and controlled vocabulary in an 'ideal' combination."

In 1980 two studies similar to Henzler's were reported in the United States. One, by Markey, Atherton, and Newton, looked at 165 free-text search statements used in accessing the ERIC database to determine whether the concepts expressed by the free-text terms could also be expressed by ERIC descriptors¹⁸. They found that one out of every eight search statements could not be represented in the controlled vocabulary. Natural-language terms not likely to be represented in the ERIC controlled vocabulary were those designating geographic areas, recent topics, specific named objects, value

judgments, actions, and individual or psychological characteristics. For six search topics, retrieval was performed using both free-text and a controlled-vocabulary formulation of the topics. Overall, free-text retrieval produced higher recall and lower precision than retrieval using a controlled vocabulary. This finding is surprising, as it is at odds with those of previous studies. It is discussed further below. The second 1980 study incorporated cost considerations. Carried out by Calkins on the COMPENDEX and ENVIROLINE databases¹⁹, it produced as one of its findings that "a limited number of citations, usually the best ones, are retrieved by using controlled indexing terms." Calkins advised that persons who want only a few pertinent items use the controlled-vocabulary terms. She went on to observe that her case study "disproves the hypothesis that a free-text strategy retrieves ALL the relevant information. Only by using a combination of the two best of terms was the maximum retrieval obtained in the database [sic] studied." In the Calkins study, search with a controlled vocabulary proved to be faster and expensive in connect time. The figures for the former give 4.5 minutes and \$5.00, as compared to 21 minutes and \$23.00 for the latter.

Other studies of a similar vein but using full-text data terms followed these two²⁰, the most significant of which was Carol Tenopir's doctoral work²¹. Using the Harvard Business Review Online database, Tenopir separately searched the full text, abstract, title, and descriptors of articles for each of 36 queries. Not surprising was the result that each of the different searches produced unique documents, and that the full text searches produced better recall and worse precision than the controlled-vocabulary search. A reason given why the fulltext method was able to extract unique documents from the database is that the vocabulary provided by the full text of a document is larger than that of any of its surrogates, i.e., its title, abstract, or descriptors; thus, this vocabulary expresses concepts not expressed by the surrogates, including more specific concepts. A second reason given for the performance of the full-text method in retrieving unique documents is somewhat worrisome. It would seem that on the database searched the full text used more synonyms than the controlled vocabulary. This is puzzling: What is a controlled vocabulary for? One is tempted to speculate that the controlled vocabulary used might not have been of the best sort.

Era Three studies were case studies and did not aim at the level of generalization aspired to by large-scale efforts like the Cranfield experiments. Consequently, little criticism was levelled against them. As research methodologies, case studies can be legitimately used to disprove hypotheses, as in the Calkins study where the hypothesis that free-text searching retrieves all relevant information was disproved. Moreover, while a single case study in itself is limited as to its generalizability, the fact that many similar studies produce the same finding lends credence to that finding. Out of the Era Three studies emerges one consistent finding, viz., that controlled and uncontrolled vocabularies have different properties, and thus behave differently in retrieval.

RESEARCH DIRECTIONS

Controlled Vocabularies and Precision:

A signal that research is needed is when research studies produce contradictory findings. The finding in the Markey, Atherton, and Newton study that free-text retrieval produced higher recall and lower precision than retrieval using a controlled vocabulary not only conflicts with results of earlier studies; it runs counter to "conventional wisdom" as well. Conventional wisdom holds that free-text terms contribute to precision by virtue of being more specific and more current than controlled vocabulary terms²². It also holds that a controlled vocabulary, by virtue of its classing functions, serves primarily to promote recall. That this is the case is elegantly argued by Fugmann using a conceptual correlate of recall called "representational predictability"²³. Fugmann argues that "the success of any search depends on the predictability of modes of expression for topics of interest." For a search on a topic, like Insects, of which there are over a million different kinds, it would be practically impossible to formulate a free-text search strategy. (Imagine an "OR" string of a million disjuncts!) However, if the term Insects were assigned to every article that dealt with any of these million kinds, representational predictability would be high and also, presumably, would recall. Fugmann uses this example to demonstrate the effectiveness of controlled vocabularies.

There is, then, a discrepancy between the Markey, Atherton, and Newton finding and the commonly held belief that controlled vocabularies promote recall at the expense of precision. How is the discrepancy to be explained? How did it come about that controlled-vocabulary searching produced better precision than searching with free-text terms? An answer to this question might in by looking at what it is that controlled vocabularies do to promote precision. Providing qualifiers for homonyms is a function provided by most thesauri and, as noted earlier, this is a precision device. However, it seems unlikely that homonym disambiguation occurs frequently enough to make the kind of impact on precision reported in the Markey, Atherton, and Newton study.

Another possible explanation for the Markey, Atherton, and Newton finding, which associates controlled vocabularies with precision, may lie in the fact that normally there are fewer occurrences of controlled-vocabulary terms in the bibliographical record for a document than free-text terms. This would be particularly true where the bibliographical record includes abstracts. Empirical and theoretical studies have shown that as indexing depth is increased, recall improves and precision deteriorates²⁴, the implication of which is that searching on a few terms in a controlled vocabulary field would produce better precision than searching on all the terms that occur in any field; in fact, precision would be comparably better and recall comparably worse, in proportion to the amount of free text that is searched. But this explanation is also suspect. A recent evaluation by Blair and Maron of the effectiveness of free-text searching a large amount of text produced quite high precision (70%)²⁵. It is possible that earlier studies that associated increasing depth with deteriorating precision were constrained by unsophisticated search methodologies, and that repeating such studies today would produce different results. In any case, research is needed to address the common persisting belief that more is better when assigning index terms or providing searchable text.

The question as to why searching using controlled-vocabulary terms may produce better precision than searching with free-text terms is still unanswered. Another possible explanation is that normally, controlled-vocabulary terms are assigned to documents by human indexers and, since assigning terms is purposeful to retrieval and involves intellectual effort, it may be hypothesized that an assigned term is a better predictor of a document's relevance than a term that occurs in a descriptor field such as a title or an abstract. Lack of operational definition of "relevance" or "aboutness" is an obstacle to be overcome in testing whether the relationship between an index term and a document is more immediate than the relationship between the same document and a term chosen at random from its title or abstract. However, some related work has been done. Several studies have shown a strong correlation between assigned index terms and title terms; other studies have produced contrary results²⁶.

Some distinctions of importance might be noted here. The issue of assigned versus derived indexing is often intertwined with that of free text versus controlled vocabularies. This is unfortunate in two respects. First, an empirical, but not necessary, correlation exists between the use of controlled vocabularies and assignment indexing. It is not inconceivable that assignment indexing with and without the use of a controlled vocabulary would produce comparable results. Considering the extensive intellectual effort that goes into constructing controlled vocabularies, this seems an obvious target for research. Second, comparisons of free-text and controlled-vocabulary searching can be obfuscated if retrieval failures attributable to vocabulary are not distinguished from those caused by the manner in which the vocabulary is used. It was mentioned earlier that a good tool can be used clumsily. *Mutatis mutandi*, free-text searching can produce undesirable results not because the text being searched is wanting, but simply because the searcher is unskilled. In either case, difficulties in controlling for human factors have led to dubious findings in research designed to compare free-text and controlled-vocabulary searching, for human factors can obfuscate comparisons of free-text and controlled-vocabulary searching.

This section began by speculating on the "why" of an experimental finding that controlled-vocabulary searching produced better precision than free-text searching. Some possible explanations were given, but others may well be conceived. In any case, controlled vocabularies can promote precision; they can also promote recall. This would seem to contradict the tradeoff hypothesis that pits precision against recall. More likely, however, is that important distinctions are not being drawn. Free-text and controlled-vocabulary terms each contribute to precision and each to recall, but they do so in different

ways and it is the relative weight of the contributions that affects any given retrieval outcome. The determinants of precision and recall cannot be simplistically conceived. Theoretical and analytical study is needed to understand the complex casual mechanisms involved.

FORMS OF CONTROL

Global comparisons of free-text and controlled-vocabulary searching tend to produce muddled results. A clearer picture might be obtained by looking at the degree and kind of control embodied in an index language and testing the effect of these on retrieval performance.

It is up to the designer of a controlled vocabulary to decide just how much control and what forms of control to incorporate in it. For instance, he may decide to have a thesaurus in which only synonyms are controlled, or one in which there is singular-plural control, orthographic control, and synonym control, but no hierarchical or related-term control. Traditionally there has been little imagination or innovation brought to design considerations of this sort; normal practice has been to design controlled vocabularies on models embodied in one of the standards for thesaurus construction. (It is the business of research to challenge established practice.) To be sure, some research has been done on evaluating the effectiveness of different forms of control, notably the Cranfield II and the Aberystwyth tests²⁷. However, it is questionable how valid the findings of these experiments are in the context of modern online systems with sophisticated searching techniques. Also, it is questionable whether retrieval experiments are the best approach to evaluating the effectiveness of different forms of vocabulary control.

A factor of critical importance in any study of the effectiveness of different degrees of control is the discipline whose vocabulary is being controlled. For instance, the effectiveness of synonym control might be expected to vary from discipline to discipline, depending upon what Bhattacharyya calls the terminological consistency of a discipline²⁸. The terminological consistency of an individual concept is the stability of the relationship between the concept and the terms referring to it. It is defined

$$t_c = c/s,$$

where c , the concept, has the value 1, s is the number of terms referring to c , and t_c is the terminological consistency of concept c . An overall measure of terminological consistency for a discipline is obtained by averaging the terminological consistency of the individual concepts. Such a measure poses problems for implementation; a method must be devised for enumerating or sampling all concepts in a field and a criterion defined for establishing when two concepts are synonymous.

A given concept may be represented by synonymous expressions; it can also be represented by expressions at different levels of generality, e.g., Mosquito and Insect. Fugmann's representational predictability might be regarded as an extension of Bhattacharyya's notion of terminological consistency in that it provides for hierarchical structurings in addition to synonym relationships. It too might be operationalized and vocabularies of different disciplines assessed as to the degree of representational predictability they exhibit.

Disciplines vary as to the terminological consistency and representational predictability of their vocabularies; the effectiveness of different forms of vocabulary control might be expected to vary accordingly. There is some empirical evidence to suggest that the results of evaluating homonym control in one discipline cannot be generalized in another. Synonym control for the discipline of aerodynamics (Cranfield II) appeared to be more effective, that is to make a bigger difference, than for the discipline of information science (Aberystwyth)²⁹. The implication is that the terminology of aerodynamics is more terminologically inconsistent than that of information science. (Eventually, the methodologies of the two experiments were not really comparable, so it would be incautious to draw any implications.)

It seems plausible that one should be able to evaluate the effectiveness of a given form of vocabulary control without bringing in the heavy artillery of a retrieval experiment. Simply examining the vocabulary of a discipline for the incidence of homonyms, synonyms, and classificatory concepts in it would give some idea of the effectiveness of control at these different levels. For instance, since hierarchical structuring is indigenous to the science, one might expect scientific texts to exhibit a great deal of representational unpredictability; thus, broader-term/narrower-term searching would be effective in searching such texts. This type of searching would not be so effective in the humanities, where the terminology employed often does not lend itself to classification. Scientific language is fairly precise in comparison with the language of the social sciences, which is notoriously rife with homonyms and polysemes; the term *Culture* has over 160 different meanings; thus, one would expect homonym control to be more effective in the social sciences than the sciences.

Some examination of the vocabularies of disciplines with a view to the effectiveness of vocabulary control has been done. Some recent Russian work indicated that the more technical the vocabulary of a discipline, the fewer synonyms it contains³⁰ (does this mean aerodynamics is more terminologically consistent than information science?) Wiberley, studying the entry terms in leading encyclopedias and dictionaries in the humanities, found that only two-fifths of these were "imprecise"³¹. Svenonius reported difficulties in attempting to structure vocabulary in the discipline of art history³². It would seem that a study of the vocabulary of a discipline should precede any attempt to control that vocabulary for information retrieval. As observed earlier, most thesauri are developed after a common pattern as established by one of the standards organizations; but would it not make more sense to custom tailor a vocabulary-control tool to the vocabulary being controlled?

AUTOMATIC VERSUS INTELLECTUAL CONTROL

How much human effort is needed in the constructing of vocabulary-control devices and how much can be provided by machine algorithms? For instance, is it necessary to relegate to a thesaurus the control of orthographic variations such as English and American spelling variants? Insofar as such variations are systematic in nature, a program can be written that will automatically equate *Color* and *Colour*, *Rationalization* and *Rationalisation*, etc. Also a program can be written to deal with words that sometimes occur in hyphenated form and sometimes not. To the extent that a vocabulary can be controlled, in the sense of classing certain kinds of like terms together, algorithmically, there is no need to spend human labor on the task. The distinction between thesaurus-imposed vocabulary control and vocabulary control that can be imposed systemically is a distinction between a thesaurus whose construction requires intellectual effort and one that does not.

This is the case for orthographic variations, the normalization of the singular and plural terms does not, for the most part, require the construction of a thesaurus. Apart from irregular forms, e.g., *Mice* and *Mouse*, singular-plural control can be implemented either by algorithm or by the user at the time of search. Nor is a thesaurus needed to precollocate terms that share common word stems or terms that are syntactic variants, e.g. the direct and inverted forms of a compound term. Again, it would be a waste of intellectual effort to link terms where the linking can be described by an algorithm either built into the system or easily constructed by a user at the time of his search.

Algorithmic methods can be used also to establish related-term relationships. The definition of two related terms is if when thinking of one of them the searcher would like to be reminded of the other. The subjectivity implied by this definition would seem to militate against algorithmic methods. On the other hand (another question for research), it may be that related-term relationships derived according to some rule might really be more effective in retrieval than those created through a subjective or scatter-shot approach.

In developing algorithms for constructing related-term relationships, co-occurrence data can be used: Either the co-occurrence of terms in texts to be searched or their co-occurrence in users' requests as embodied in transaction logs. Terms that co-occur frequently in texts might be expected to have something in common, i.e., to be related. In the AID system, an operational system that makes use of

co-occurring terms, a search on Shellfish will yield in descending order of relatedness: Oysters, Mussels, Clams, Tides, Estuaries, Parahaemolyticus, Crassostrea, Seafoods, and Virginia²².

Alternatively, the construction of related-term relationships can be based on co-occurrence data found in transaction analyses. The use of transaction logs has three reasons to recommend it. First, the semantic associations attaching to a given term may be very large; however, they could well be significantly reduced when the term is used for the purpose of information retrieval. Second, a corollary of Bradford's law is that past use is a predictor of future use; thus, the fact that terms have been associated in search strategies in the past increases the probability that they will be associated in future searching. Third, certain thesaurus standards, e.g., the ISO and UNISIST standards, legislate that related-term relationships should be established only if they will be required in retrieval; co-occurrence in past search strategies would count as demonstrable evidence or "use warrant" for meeting this criteria. Some online systems already make available to their users data extracted from transaction-log analyses²⁴. A record is kept of terms have been correlated in past search strategies in the form of Boolean disjunctions. Terms frequently strung together in "OR" searches, called "hedges," are then made accessible to current users of the system in the expectation that they will be of use in formulating search strategies.

Algorithmic methods used to derive related terms may serve as well for the derivation of synonymous terms. Definitionally the two types of terms may be regarded as forming a continuum. Similar arguments favor the use of such methods. Particularly persuasive is the Bradford-based argument that most of the use of synonym relationships in thesauri is concentrated in a few such relationships. This argument is a deductive one deriving from a generalization of Bradford's law to all objects produced for human use. However, we really do not know how many of the synonym linkings in a thesaurus are actually utilized in a retrieval system; Empirical study is needed to ascertain this and to determine if patterns of synonym use would warrant the generating of synonym linkings not *a priori* but *a posteriori* on the basis of past use as recorded in transaction log analyses.

It would appear that hierarchical (broader-term/narrower-term) relationships cannot be detected automatically, at least not at present. Strictly interpreted, the hierarchical relationship is the transitive, reflexive, and asymmetrical relationship logically defined as set inclusion. To determine correctly when a hierarchical relationship holds requires not only logical sense but also real-world knowledge. For instance, it must be possible to assign the truth value "true" to the statement "all parrots are birds" and the value "false" to "all parrots are pets." Conceivably a machine could make such truth value assignments, if embedded in it were knowledge representations embodying classificatory structures. But this does not obviate the need for intellectual input; it just removes it one step back. Someone might at this point argue that methods of automatic classification indeed exist and these might be used to automatically generate classificatory structures. However, these classifications are polythecal, rather than hierarchical in nature, and their proper use is in contributing to related-term structurings.

USER-IMPOSED VOCABULARY CONTROL

The possibility of algorithmic methods of vocabulary control is attractive in the promise it offers of intellectual relief and cost savings in thesaurus construction. Where algorithmic measures cannot be used to effect vocabulary control, intellectual effort is required; this involves expense, either to indexers constructing controlled vocabularies or users attempting to search without them. The question can be raised whether it is not reasonable to expect users at the time of search to provide a certain amount of control? One argument for *not* providing users with extensive vocabulary control, but rather expecting them to impose it themselves, is that many thesaurus relationships are viewpoint dependent. This is particularly true of related-term and synonymous relationships. For instance in designing a search, one user might wish to regard Salons, Parlors, and Living rooms as synonymous. For his purposes, the differences between them can be regarded as indistinguishable. Another user, however, with different purposes, might insist the distinctions be recognized in retrieval. The fact that it would be difficult to construct a thesaurus that incorporated all possible viewpoints is an argument for user-imposed control, at least for seldom-used synonymous and related-term relationships. The extent of the divergence of viewpoints among users is a matter worth some study.

On the other hand, a persuasive reason for *not* relegating vocabulary-control functions to users at the time of search is that users do not have the verbal imagination to conceive of all possible search terms. One of the functions of a controlled vocabulary is to stimulate users' verbal imaginations, not only in suggesting possible synonyms but also in helping them to express their search concepts at the desired level of generality. It has been several times demonstrated, at least in manual environments using the *Library of Congress Subject Headings*, that the terms chosen by users to express concepts often are broader or narrower than those used by the system³⁵. Users' verbal abilities might be studied more abstractly; to what extent can users predict how given concepts will be represented in documents: at what levels of generality is a given concept expressible and in what equivalent guises? Given that some vocabulary control is needed, it is important to ask how much can be system, i.e., algorithmically, supplied; how much can be supplied by users themselves; how much must be supplied in the form of a controlled vocabulary such as a thesaurus?

Another argument for not relegating vocabulary control to users at the time of search is that they do not have time to key in all possible search terms, e.g., the Fugmann example where for a search on insects the names of every insect existent must be entered into a search request. As suggested earlier this is an argument that probably holds more sway for scientific texts than texts in the social sciences or humanities. However, in any case, the general question is important: How much vocabulary control can be supplied by users themselves and how much must be supplied in the form of a controlled vocabulary? The point is sometimes made that the time that indexers put into constructing a controlled vocabulary for a system is offset by the time saved by those using it. While this might be a truism, some value would be gained in seeing it demonstrated with actual payoff figures. Factors to be taken into account in such a demonstration include users' retrieval requirements, expressed in terms of precision and recall, as well as the incidence and seriousness to users of different types of failures: when too little is retrieved, when too much is retrieved, and when not everything is retrieved on a given subject.

SUMMARY

After reviewing briefly the history of the issue of free-text searching versus searching with a controlled vocabulary, this paper has suggested areas for research that would clarify the issue in such a way as to contribute to a rational basis for the design of retrieval tools such as thesauri. Among questions in need of research are the following:

- What is the retrieval effectiveness of human and/or machine indexing with and without a controlled vocabulary? An answer to this question would serve to distinguish the relative impacts on retrieval of the vocabulary used and the fact that it is assigned.
- What is the effect of depth of indexing on retrieval effectiveness?
- What is the effect of the amount and kind of text searched on retrieval effectiveness?
- What is the effect of different kinds or degrees of vocabulary control on retrieval effectiveness?

The last three questions have, of course, been asked before, and research has been addressed to them; however, it seems appropriate to raise them again and ask for answers designed more carefully to take into account the variables of search methodology and the discipline searched.

To be classed more as development than research is the devising of ever more sophisticated search methodologies incorporating procedures devised by information scientists studying different methods of automatic indexing and by linguists studying different methods of text analysis. Once developed, however, these search methodologies themselves become a target for evaluation. Thus further research questions:

- What is the effect on retrieval performance of automating different forms and kinds of vocabulary control?
- Given that some form of control is needed, what is its proper locus: the system, the "controlled vocabulary," or the user?

As observed earlier, questions of effectiveness are inevitably associated with questions of cost. Thus, the question:

- What are the relative costs of different approaches to vocabulary control?

In the past, answers to questions of effectiveness have been approached primarily through retrieval experiments, which have produced equivocal results. But other approaches may be possible. It is, after all, the business of research to overcome obstacles that stand in the way of finding answers; it may be that a way can be found to assess retrieval effectiveness without incurring the expense and control problems involved in large-scale retrieval experiments, yet at the same time achieving a generality beyond what is possible through case studies. Some questions that can be researched outside the context of large-scale or case-study retrieval experiments are:

- What is the ability of users to verbalize search requests, particularly in imagining alternative formulations of their search concepts, but also in recognizing and disambiguating homonyms?
- What is the relevance or aboutness to an article of different kinds of terms: those assigned by indexers, those found in the titles of the articles, those found in the abstracts, and those distributed in different ways in the article itself. Again, work has been done in this area, but, it seems safe to say, it needs to be complemented by an agreed-upon, operational definition of *aboutness*.
- What is the semantic structure of text in a given discipline, particularly as to the degree of homonymy it embodies and the degree to which concepts can be formulated by synonymous terms, at different hierarchical levels and from different points of view? Probably this is the single most critical question that needs to be addressed before undertaking the design of a controlled vocabulary.

REFERENCES

1. Quoted in Cutter, Charles A. "Library Catalogues." In: *Public Libraries in the United States of America: Their History, Condition and Management*. Special Report. U.S. Bureau of Education. Washington, D.C.: U.S. Government Printing Office; 1876: 535.
2. *Ibid.*, p. 531.
3. *Ibid.*, p. 535.
4. Crestadoro discusses his method in Crestadoro, Andrea. *The Art of Making Catalogues of Libraries*. London: The Literary, Scientific & Artistic Reference Office. 1856.
5. Metcalfe, John. *Information Indexing and Subject Cataloging: Alphabetical: Classified: Coordinate: Mechanical*. New York: Scarecrow Press: 1957: 47.
6. Cutter, *op. cit.*, p. 536.
7. Luhn, Hans Peter. IBM Technical Report No. RC-127. Yorktown Heights, NY: IBM; 1959.
8. There were two Cranfield experiments: (1) Cleverdon, Cyril W. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield, England: College of Aeronautics, ASLIB Cranfield Research Project; October, 1962. (2) Cleverdon, Cyril W.; Mills, Jack; Keen, Michael. *Factors Determining the Performance of Indexing Systems*. Vol. 1, *Design*, Parts 1 and 2; Vol. 2, *Test Results*. Cranfield, England: College of Aeronautics, ASLIB Cranfield Research Project; 1966.
9. For instance, Parker, J. E. "Preliminary Assessment of the Comparative Efficiencies of an SDI System Using Controlled or Natural Language for Retrieval." *Program*. 5:26-34: 1971. Aitchison, T. M.; Hall, A. M. "Evaluation of Retrieval Effectiveness." In: L. Vilentchuk. Ed. *Proceedings of the ISLIC International Conference on Information Science, Tel Aviv, August 29 - September 3, 1971*. Tel Aviv: The National Center of Scientific and Technological Information: 1972: 373-383. Keen, E. Michael. "The Aberystwyth Index Languages Test." *Journal of Documentation* 29: 1-35; 1973.
10. Swanson, Don R. "Evidence Underlying the Cranfield Results." *Library Quarterly*. 35:1-20: 1965.
11. Svenonius, Elaine. "Good Indexing: A Question of Evidence." *Library Science with a Slant to Documentation*. 12:33-39; 1975.

BEST COPY AVAILABLE

12. Keen, E. Michael, *op. cit.*, p. 33.
13. Svenonius, Elaine. "Natural Language vs. Controlled Vocabulary." *Proceedings of the Fourth Canadian Conference on Information Science, London, Ontario, May 11-14, 1976*. Ottawa: Canadian Association of Information Science: 1976: 141-150.
14. Charton Barbara "Searching the Literature for Concepts." *Journal of Chemical Information and Computer Sciences* 17:45-46: 1977.
15. Rowlett, Russell J. Jr. "Keywords vs. Index Terms." *Journal of Chemical Information and Computer Sciences*. 17:192-193: 1977.
16. Carrow, Deborah; Nugent, Joan. "Comparison of Free-Text and Index Search Abilities in an Operating Information System." *Information Management in the 1980s: Proceedings of the American Society for Information Science 40th Annual Meeting, September 26- October 1, 1977*. White Plains, NY: Knowledge Industry Publications; 1981: 131-138.
17. Henzler, Rolf C. "Free or Controlled Vocabulary: Some Statistical User-Oriented Evaluations of Biomedical Information Systems." *International Classification*. 5(1):21-26: 1978.
18. Markey, Karen; Atherton, Pauline; Newton, Claudia. "An Analysis of Controlled Vocabulary and Free Text Search Statements in Online Searches" *Online Review* 4:225-236; 1982.
19. Calkins, Mary L. "Free Text or Controlled Vocabulary? A Case History Step-By-Step Analysis ... Plus Other Aspects of Search Strategy." *Database*. 3:53-67: 1980.
20. For instance, Durkin, Kay *et al.* "An Experiment to Study the Online Use of a Full-Text Primary Journal Database." In: *Proceedings of the 4th International Online Information Meeting, London, December 1980*. Oxford, England: Learned Information Ltd.: 1980: 53-56. See also Hersey, David F. *et al.* "Free Text Word Retrieval and Scientist Indexing Performance Profiles and Costs." *Journal of Documentation* 27: 1967- 1983; 1971.
21. Tenopir, Carol. "Full Text Database Retrieval Performance." *Online Review* 2:149-164; 1985.
22. For instance, Lancaster, F. Wilf. "Vocabulary Control for OnLine Interactive Retrieval Systems: Requirements and Possible Approaches" In: A. Neelameghan, Ed. *Ordering Systems for Global Information Networks: Proceedings of the Third International Study Conference on Classification Research, Bombay, January 1975*. Bangalore, India: International Federation of Documentation, Committee on Classification Research: 1979: 40-53 See also Norris, C. "MeSH - the Subject Heading Approach." *ASLIB Proceedings*. 33:153-159; 1981.
23. Fugmann, Robert. "The Complementarity of Nature and Indexing Languages. *International Classification* 9:140-144; 1982.
24. Empirical demonstration was given in the Cranfield studies. *op. cit.*, Theoretical demonstration was given in Swanson, Don R. "On Indexing Depth and Retrieval Effectiveness." In: *Second Congress on the Information Systems Sciences*. Washington, D.C.: Spartan Books. 1965: 311-319.
25. Blair, David C.; Maron, M. E. "An Evaluation of Retrieval Effectiveness (or a Full-Text Document-Retrieval System)." *Communications of the ACM*. 28:289-299: 1985.
26. For instance, Tell, Bjorn V. "Retrieval Efficiency from Titles and the Cost of Indexing." *Information Storage and Retrieval*. 7:241-243; 1971. Henzler, Rolf G., *op. cit.*, Bloomfield, M. Simulated Machine Indexing. Part 2. Use of Words from Title and Abstract for Matching Thesauri Headings." *Special Libraries*. 57:232-235: 1966. Montgomery, Christine; Swanson, Dan R. "Machine Indexing by People." *American Documentation*. 13:359-366; 1962.
27. Keen, E. Michael, *op. cit.*
28. Bhattacharya, K. "The Effectiveness of Natural Language in Science Indexing and Retrieval." *The Journal of Documentation* 30:235-254: 1974.
29. *Ibid.*, p. 236.
30. Andrykovick, P. F.; Korolev, E. I. "The Statistical and Lexico-grammatical Properties of Words." *Automatic Documentation and Mathematical Linguistics* 11:1-11; 1977.
31. Wiberley, Stephen E. Jr. "Subject Access in the Humanities and the Precision of the Humanist's Vocabulary." *Library Quarterly*. 53:420-433; 1983.
32. Svenonius, Elaine. "Information Retrieval in the Field of Art." In: L. Corti and M. Schmitt, Eds. *Proceedings of the Second International Conference on Automatic Processing of Art History Data and Documents, held at the Scuola Normale Superiore, Pisa, September, 24-27, 1984*. Pisa: Scuola Normale, Regione Toscana, 1984: 33-48
33. Doszkocs, T. E. "An Associative Interactive Dictionary for Online Searching." *Online Review* 2: 168-173: 1978.

34. Sievert, Mary Ellen and Boyce, Bert R. "Hedge Trimming and the Resurrection of the Controlled Vocabulary in Online Searching." *Online Review*. 7:489-494: 1983.
35. For instance, Bates, Marcia. "System Meets User: Problems in Matching Subject Search Teams." *Information Processing and Management*. 13:367-368: 1977.

SVENONIUS: This paper was first published in the *Journal of the American Society for Information Science: JASIS* 37(5) 331-340, 1986 (ISSN 0002-8231) and has been reproduced here with the kind permission of John Wiley & Son, Inc.

NEEDS FOR RESEARCH IN INDEXING

Jessica L. Milstead

The JELEM Company
P.O. Box 5063
Brookfield, CT 06804

In recent years, the amount of research in indexing appears to have decreased, despite the continued need for improvement in both quality and cost-effectiveness of indexing. The primary purpose of any index is to permit users to locate information; this implies both a need for research into user needs *per se*, and a requirement that any research be oriented toward the goal of meeting those needs. Cognitive processes of both indexers and index users need study; the former appear not to have been explicitly addressed in the literature. Issues of vocabulary control - the extent of control, or even whether it should be used at all - continue to present problems, despite the continued development of tools intended to aid in such control. Increased computer power has made it possible to relegate to the computer more of the tasks formerly carried out by humans. The need to determine how best to supplement human intellectual effort with computer capabilities has become urgent. Structure and layout of indexes, both on the printed page and on computer screens, remain primarily a matter of folklore and intuition, despite the existence of some research on the topic. Finally, more research on evaluation of indexes is required, to assure that all other efforts contribute to actual usability of the final product.

Over the years, the amount of research in indexing - at least as reported in the *Annual Review of Information Science and Technology* (ARIST) - seems to have decreased significantly. In the first 10 years of ARIST (1966-1975), there were an average of 25-30 index entries per year under "Indexing" and its topical cross-references. Since then, the average has been closer to 10-12 entries per year. Yet, there is little evidence that many of the problems in either the making or the use of indexes have been resolved.

While it is impossible to determine precisely why the amount of research on indexing decreased, some reasonable guesses may be made. Relatively little of the research was actually applied in practice. With the exception of Keen's work (1973), much of it was applicable primarily to databases, not page-layout print products. As technology changed, it became possible to work with much larger databases, even with full text. As a result, researchers became more interested in work on natural language retrieval, avoiding indexing entirely. Thus, the trend in recent years has been to work on methods of retrieval from natural language text, whether brief documents such as abstracts or full narratives, rather than on indexing (i.e., analysis of individual documents at input) as such.

Two countertrends have been evident recently. The first is in research and systems designed to categorize documents automatically. This work, however, has limited application to indexing in the normal sense, because such systems typically are limited to specified types of documents; they may not attempt to categorize all documents in a system, but only those that meet certain criteria. The system in use to classify Reuters newswire stories is an example (Hayes & Koerner, 1993).

Indexes continue to be both compiled and used by humans, and there is no evidence that this situation will change in the near future. While databases that have no human indexing are available, there is even a tendency for some human effort to be added to these. Indexes in print will continue to exist for at least

as long as paper books exist, and no amount of natural language retrieval is going to aid in finding information accessible only in paper. Therefore, indexing research should not be limited to indexes maintained and searched in electronic media.

The second countertrend reflects this need. After a near-total dearth of research on print indexes since Keen's work (1973), there have recently been several studies that describe the characteristics of print indexes (Bishop, Liddy, & Settel, 1991; Diodato, 1991; Diodato & Gandt, 1991; Liddy, Bishop, & Settel, 1991). There has even been at least one preliminary report of a test of search of print indexes (Liddy & Jorgensen, 1993).

Indexes are designed for people to use. This should be self-evident - but perhaps it is too self-evident. Indexers need answers to a fundamental set of questions about users, particularly questions about their cognitive state, and the amount of effort that they can be expected to undertake and under what circumstances. It is important to realize that there is no such thing as a generic "user," and that the answer to a given question will vary by type of user. This simply adds to the complexity and urgency of the research needs, because we need to know the answers for different groups of users.

Therefore, it is unfortunate that more research is not being carried out in indexing, and it is worthwhile to look at some questions on which research is needed. The needs for research may be divided into five areas:

- cognitive aspects;
- vocabulary control;
- machine-aided indexing;
- layout and structure; and
- evaluation.

Shaw and Fouchereaux (1993) have listed areas of needed research, drawn from the last few years of ARIST. Bakewell (1993) also summarizes indexing research needs.

Kleinberg (1993) reviewed five studies of back-of-book indexes. While the findings of these studies were all enlightening, they were essentially descriptive, providing no guidance in determining how well indexes meet user needs.

Babu (1992), on the other hand, surveyed philosophy scholars and teachers with regard to their preferences in indexes. While the study is limited, it appears to show that this group of users - philosophers in India - is very aware of indexes and has some definite preferences for their contents and arrangement.

Diodato (1994) conducted a brief survey of the preferences of librarians and college professors, asking them specific questions about alphabetization, format, and use of cross-references. While it would be desirable to confirm his results with more in-depth research, his findings are illuminating.

The work presently underway on revision of the ANSI/NISO standard for indexes points up the need for research. As has always been the case with standards for the content of index and catalog records, this standard is being developed by a group of experts who, with input from the field, are trying to codify good practice. Resource constraints have prevented any research from being undertaken so that the standard will be based on these experts' considered judgment, together with extensive input from the field. Yet it seems certain that the standard could be improved if its provisions were to be based on solid research findings.

This article is concerned with all indexes for information retrieval, in whatever format they may be available. The ability of actual indexes to facilitate information retrieval in real-life situations is the point. What might be theoretically possible is of less interest.

COGNITIVE ASPECTS

Cognitive Processes in Indexing

Perhaps the most important need for research is one that appears never to have been directly addressed. This is also Shaw and Fouchereaux's (1993) first need: "What are the cognitive processes involved in indexing and classification?" The first two basic processes in indexing are deciding what a text or information item is "about," and/or what questions it will answer; and translating that decision into the terms used in the indexing system - whether free text or controlled vocabulary. The second of these steps can be described with some degree of accuracy, but we have no idea of the mental processes involved when an indexer decides what a piece of information is "about," or what user queries it is likely to satisfy.

This gap in our knowledge is critical from two points of view. First, it has meant that automatic or computer-aided indexing systems can not produce indexing comparable to that produced by humans. They must operate in fundamentally different ways, and typically require extensive interaction with the searcher to produce results comparable to those available by search of a more conventional system. Present-day decision rule systems, as well as statistical, syntactic, and semantic methods, attempt to mimic the results produced by humans, but their success is limited because of our lack of understanding of the processes which are actually involved.

Second, the known high level of inconsistency in indexing must arise at least partly from differences in perception either of the "aboutness" of the document or of the value of a particular piece of information to users of the database. Any case in which two indexers analyzing the same document apply terms that denote quite different concepts must arise from such differences in perception. In fact, it seems reasonable to assume that once differences in training, subject expertise, and familiarity with the vocabulary and indexing policies are accounted for, what remains are differences in cognitive experiences with the document.

Cognitive Processes in Searching

Complementing the need for research on indexers' cognitive processes is that for study of those of index users (Allen, 1991). Much research has been conducted on information needs and use and on information retrieval, but less on usability. How do users frame their queries? How can indexes assist users in articulating their needs so that they can locate the information that the document or collection has to offer?

Another of Shaw and Fouchereaux's three research needs speaks to this issue: "How can machine-readable data be organized to improve perception of relationships and retrieval?" (Preschel & Woods, 1989).

Organization of indexes to fit users' cognitive processes includes such issues as layout and structure, arrangement, and the extent and manner in which related concepts are brought together. These issues are discussed in later sections of this article.

VOCABULARY CONTROL

Thesaurus Relationships

Do users who actually refer to a thesaurus know or care about the meaning of different kinds of relationships (BT/NT, RT, USE/UF), or are these of value exclusively for thesaurus and index designers and maintainers? Even if the latter is the case, it still may be worthwhile to maintain such relationships, because of their value in both indexing and vocabulary maintenance. It is worthwhile knowing, however, whether we are maintaining such relationships solely for our own use or for the use of at least a subset of our users.

A considerable body of research, conducted over the past quarter century, has been interpreted to indicate that various natural language devices can substitute for human indexing using a controlled vocabulary. Up to the present time, however, it has not been found practical in database indexing to abandon use of controlled vocabularies. This is not entirely due to conservatism. Rather, databases must be indexed to suit the capabilities of the systems used to search them - and these systems have historically been limited essentially to Boolean and proximity searching. Even as access to full text has become available, indexing has remained necessary because the capabilities of retrieval systems to sort sets on the basis of importance of the items retrieved have been minimal at best.

While this situation is beginning to change, it will be some time before full-text retrieval engines will be powerful enough to substitute fully for human indexing - if this ever happens. Furthermore, all indications are that even when such retrieval substitutes for human indexing, its results are improved by use of vocabulary management tools or knowledge bases. Therefore, research on controlled vocabularies will continue to be needed for the foreseeable future.

Size of Thesauri

How large should a thesaurus be, in relation to the database to which it is applied? Is there some optimum that provides the best achievable balance between usability of the thesaurus and precision of retrieval? At what point does a thesaurus become so large that its use causes documents on similar topics to be scattered under different terms?

Is there a certain maximum size beyond which a thesaurus becomes less useful for both indexing and retrieval? For example, when the thesaurus reaches a certain size, are terms lost from human memory and therefore not used for either indexing or retrieval?

Balance of Vocabulary Control and Natural Language

In an electronic index, what is the best balance between natural language and vocabulary control? To what extent can reliance be placed on natural language searching for retrieval, and how much assistance should be provided by manual, controlled-vocabulary indexing? There is some research - and many anecdotes - that shows that the best retrieval situation is a mix of human indexing using a controlled vocabulary and the ability to search full text. But there appears to be no work that will assist the index designer in deciding how much expensive human effort is really warranted.

Furthermore, as natural language retrieval systems become more powerful, the balance may be expected to change. Thus, any research must take into consideration the increasing capabilities of retrieval software.

MACHINE-AIDED INDEXING AND RETRIEVAL

Candidate Term Systems

Several database producers are using machine-aided indexing software today, and more are looking at the possibilities. Typically, these packages analyze text (often abstracts) and suggest terms for assignment. It is the responsibility of the indexer to review these suggested terms, deleting the inappropriate ones and adding others as needed. Most of the small amount of research on such packages reported to date has been carried out by their developers. The systems in operation at the American Petroleum Institute (Martinez, Lucey, & Linder, 1987) and at the Defense Technical Information Center and the National Aeronautics and Space Administration (Buchan, 1987; Klingbiel, 1985) are the best-known of such systems. In recent years, Access Innovations, Inc., has also begun offering contract services with a machine-aided indexing system.

As use of these methods spreads, it is becoming urgent to determine their impact on the quality of indexing. To what extent does the fact that the system has suggested a term influence an indexer to

accept it, even if it would not have been considered useful in the manual process? Conversely, to what extent does the system suggest useful terms that would have been missed in manual indexing? Does the indexer tend to go beyond the system's recommendations to use terms that the automatic routine missed?

Over time, as indexers become more familiar with the machine-aided indexing package, do they become more demanding, producing higher quality indexing? Or do they trust the results of the system, moving on to the next document without thorough review?

What is the impact of these systems on costs in the long run? Do they save enough time in indexing to repay both development and operating costs? Do cost savings come at the expense of "speedup" pressures and possible reduced quality?

Supplementing Human Indexing

A different approach to machine-aided indexing takes initial human input and adds - automatically or as candidates - additional terms that are implied by the human work. The best-known work is that of Humphrey (1989) at the National Library of Medicine. In her MedIndEx system, a knowledge base is used to expand on and correct human indexing.

There are almost certainly other areas of knowledge in which expert knowledge of the structure of the information could be used both to assist in the indexing process and to make indexing more consistent. Every ongoing indexing operation has its policies and rules, codified to a greater or lesser extent. Many of these rules would be amenable to further systematization in an expert system, eliminating some indexing steps, and assuring more consistent following of policies.

The extent to which development of such systems would be cost effective is not known today, because not enough research has been carried out. Furthermore, as expert system technology develops, these aids will become more practical.

Combining Manual Indexing and Natural Language Retrieval

From another point of view, how can human indexing and statistical, semantic, and syntactic information retrieval techniques be combined in a cost-effective manner? These techniques have been developed to the point where there are a number of systems available on the commercial market. However, it seems to be true that they provide the best results on documents which have undergone some human analysis. It would be very useful to know which aspects of human indexing are least adequately replaced by natural language retrieval, so that indexing resources could be concentrated on these aspects to achieve synergistic results.

Categorization Systems

As noted above, there is a good bit of work going on today in organization of texts into a predefined limited set of categories. It appears that such categorization can be quite reliable. Could automatic categorization make human indexing more reliable or efficient? An expert system in use at the International Nuclear Information System implies that it can. This system is part of the quality control process: it uses a set of rules to check the category assignment of a document against its index terms. When a failure to match occurs, the item is tagged for human review. This system has superseded the former random checking of indexing, because it quite reliably picks up the documents that have been poorly indexed or categorized (Todeschini, 1993).

LAYOUT AND STRUCTURE

The needs in this section apply to print and electronic indexes in different degrees. However, many issues that seem to be print-oriented actually apply to the presentation of text on a screen as well. In

fact, the smaller amount of information that can be fitted on a computer screen means that this medium is even less forgiving in matters of layout. While many of the issues relate to matters of detail, these details greatly influence the searchability of any index of significant size. The examples given here are intended primarily to be illustrative; almost every aspect of index layout and structure would profit from solid knowledge of its impact on users.

Cross References

Do users really know what "see" and "see also" references mean, and the difference between them? What is the maximum useful number of cross references under a given term, that is, the number above which the typical user will not be motivated to follow up any of the references? In their preliminary results, Liddy and Jorgensen (1993) even found that cross references confused and frustrated some users.

In a database, how much value should be placed on making cross references readily available, and how can this best be done?

In a back-of-book type of index, that is, one with headings and modifiers, is it more useful to place a series of cross references under the main heading, or to spread them out under modifiers? For example:

Indexing. *see also* Automatic indexing: Chain indexing

or

Indexing

...
automatic. *see* Automatic indexing

...
chain. *see* Chain indexing

...

Similarly, should "see also" references precede or follow modifiers? For example:

Indexing, 32, 56. *see also* Automatic indexing
research, 183-185
training, 74

or

Indexing, 32, 56
research 183-185
training, 74
see also Automatic indexing

Arrangement

Should arrangement be word-by-word or letter-by-letter? Does the kind of tool affect the answer to this question? With either of these choices, should the kind of entry (e.g., place or topic) affect arrangement, or should arrangement be solely on the basis of the characters themselves? In all questions of arrangement, there are fervent believers on both or all sides, but little or no empirical evidence.

In systems using ad hoc modifiers, should prepositions and function words at the beginning of a modifier be regarded or ignored in arrangement (Wellisch, 1993)? For example:

Preposition at beginning of modifier ignored in arrangement:

Cross references

- between adjacent headings
- application on first survey
- blind, in index evaluation
- in KWIC concordances
- from multiword headings to inverted headings
- among synonyms
- in thesauri

All words in modifier considered in arrangement:

Cross references

- among synonyms
- application on first survey
- between adjacent headings
- blind, in index evaluation
- from multiword headings to inverted headings
- in KWIC concordances
- in thesauri

Length

How large should indexes be in relation to the length of the text? This is obviously not a simple question; the ratio of index length to text length varies immensely between a library catalog and a back-of-book index. Even in the latter, the index to a detailed monograph will be longer than that to a simpler work.

VALUATION OF INDEXING

Retrieval Evaluation

The primary means available for evaluation of the quality of indexing is the recall/precision methodology and its variants. While this methodology certainly provides useful information, it has always been known to suffer from significant defects. Two are probably the most important:

- Measuring recall in a real collection would require evaluating every document in the collection for its relevance to every query, a task that is effectively impossible. Studies which get around this problem by searching from a variety of points of view, or searching very broadly, and then combining all the relevant retrievals to provide the base of relevant documents, have no way of guessing how much might have been missed.

- Relevance as a measure is inherently defective. In practice, it almost has to be treated as an absolute - a particular document's degree of relevance to a query is unchanging. The decision is typically based on the degree of match between judges' determination of what the document is "about" and the terms in the query. Yet it is well-known that numerous factors other than such matching affect the usefulness of a document in a real-life situation. Not only must the document match the user's knowledge level, its value depends partly on what the user has already seen. That is, if two documents are similar, the first one the user sees may be very useful, and the second one of minimal value because it adds little that is new to the user's knowledge state.

Numerous other measures have been proposed, but none are really practical to use for evaluation of large, real-life systems. For instance, utility measures take into account the effect of a particular retrieval on a user's state of knowledge. However, the utility of a document will vary not just with

individual users, but as their knowledge state changes over time. Thus, it seems that a great many more measurements would be required in order to achieve results in which confidence could be placed. Froehlich (1994), in his introduction to a special issue of the *Journal of the American Society for Information Science* on relevance research, makes a similar distinction between "topical relevance" and "user relevance".

A set of measures that would permit reliable testing or indexing systems, preferably without a huge investment of resources, is urgently required. Many of the research needs discussed elsewhere in this article require such measures.

Usability Testing

Software and other products undergo usability testing today. Why not indexes? The obvious answer is that usability testing is very costly, and profit margins on index sales are not adequate to support such testing. Nonetheless, such research would be extremely useful.

Value of Indexing

Finally, the issue on which else hinges: Is indexing worth doing anyway? If all of the resources that presently go into indexing were put into some means of aiding retrieval, would information users get more total value or less? How would the value be measured?

From the opposite point of view, what is indexing worth in terms of the value received by index users? How much value does a given increment in the resources dedicated to indexing produce? Human indexing is visibly costly, and the pressure to cut these costs is never-ending. Unfortunately, we have little idea of what such indexing is "worth." How much does the index to a book increase its value, and how can this increase be measured? How much value does indexing add to a service which provides bibliographic records and abstracts or full text, and again, how can the added value be measured? These and other questions of the value of indexing go to the heart of the issue, but they may also be the most difficult to study.

CONCLUSIONS

People have been making indexes for hundreds of years, but there has been relatively little research on indexing as such. The bulk of the research that has been conducted is directed at the goal of substituting computer resources for human resources, rather than determining what users actually need from indexes.

The methods by which indexes are produced are changing, and the allocation of resources may be changing to emphasize retrieval - that is, fewer resources will be dedicated to analysis of individual documents at input and more to the output side of the equation. The new technologies that are becoming available make it possible to achieve information retrieval goals in different ways. It is unfortunate that these changes are coming without study of fundamental questions of what users really need and what the value of indexes really is.

ACKNOWLEDGMENTS

As part of the preparations for this article, opinions were gathered from the field by posting messages on two Internet listservs, index-l and libref-l, as well as the indexing forum on The Well. Numerous useful suggestions from these sources have been included and the contributions of these correspondents are gratefully acknowledged. However, full responsibility for interpretation, as well as for errors of both omission and commission, remain with the author.

REFERENCES

- Allen, B.L. (1991). Cognitive research in information science. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology* (vol. 26)(pp. 3-37) Medford, NJ: Learned Information.
- Babu, B.R. (1992). User preferences for indexes in philosophical books: A survey, *Library Review*, 41, 47-55.
- Bakewell, K.J.B. (1993, April). Research in indexing: more needed? *Indexer*, 18, 147-151.
- Bishop, A.P., Liddy, E.D., & Settel, B. (1991). Index quality study, part I: Quantitative description of back-of-the-books indexes. In *Indexing Tradition and Innovation: Proceedings of the 22nd Annual Conference of the American Society of Indexers* (pp. 15-51). Port Aransas, TX: ASI.
- Buchan, R.L. (1987). Computer-aided indexing at NASA. *Reference Librarian*, 18, 269-277.
- Diodato, V. (1991). Cross-references in back-of-book indexes. *Indexer*, 17, 178-184.
- Diodato, V. User preferences for features in back of book indexes. *Journal of the American Society for Information Science*, 45, 529-536.
- Diodato, V., & Gandt, G. (1991). Back of the book indexes and the characteristics of author and nonauthor indexing: Report of an exploratory study. *Journal of the American Society for Information Science*, 42, 341-350.
- Froehlich, T.J. (1994) Relevance reconsidered - Towards an agenda for the 21st Century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45, 124-134.
- Hayes, P.J., & Koerner, G. (1993). Intelligent text technologies and their successful use by the information industry. In M.E. Williams (Ed.), *14th National Online Meeting Proceedings* (pp. 189-196). Medford, NJ: Learned Information.
- Humphrey, S.M. (1981). MedIndEx system: Medical indexing expert system. *Information Processing and Management*, 25, 73-88.
- Keen, E.M. (1973, March). The Aberystwyth index language test. *Journal of Documentation*, 29, 1-35.
- Kleinberg, I. (1993). Making the case for professional indexers: Where is the proof? In N.C. Mulvany (Ed.), *Indexing, providing access to information: Looking back, looking ahead: Proceedings of the 25th Annual Meeting of the American Society of Indexers* (pp. 139-146). Port Aransas, TX: ASI.
- Klingbiel, P.H. (1985). Phrase structure rewrite systems in information retrieval. *Information Processing and Management*, 21, 113-126.
- Liddy, E.D., Bishop, A.P., & Settel, B. (1991). Indexing quality study, part II: Publishers' survey and qualitative assessment. In *Indexing tradition and innovation: Proceedings of the 22nd Annual Conference of the American Society of Indexers* (pp. 53-79). Port Aransas, TX: ASI.
- Liddy, E.D., & Jorgensen, C.L. (1993). Reality check! Book index characteristics that facilitate information access. In N.C. Mulvany (Ed.), *Indexing, providing access to information: Looking back, looking ahead: Proceedings of the 25th Annual Meeting of the American Society of indexers* (pp. 125-138). Port Aransas, TX: ASI.
- Martinez, C., Lucey, J., & Linder, E. (1987). An expert system for machine-aided indexing. *Journal of Chemical Information and Computer Sciences*, 27, 158-162.
- Preschel, B.M., & Woods, L.J. (1989). Social science information. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology* (vol. 24) (p. 279). Amsterdam: Elsevier.
- Shaw, D., & Fouchereaux, K. (1993). Research needs in information science. *Bulletin of the American Society for Information Science*, 19, 25.
- Todeschini, C. (1993). Personal Communication.
- Wellish, H. (1993, Jan./Feb.). Function words in subheadings. *Key Words*, 1, 8-10.

MILSTEAD: This paper was first published in *Journal of the American Society for Information Science: JASIS* 45(8) 577-582, 1994 (ISSN 0002-8231) and has been reproduced here with the kind permission of John Wiley & Son, Inc.

INTELLIGENT TEXT PROCESSING, AND INTELLIGENCE TRADECRAFT

Michael L. Weiner, Elizabeth D. Liddy

Mike Weiner, TextWise, 1100 Chase Square, Rochester, NY 14604, USA. E-mail: 74405.1451@compuserve.com
Elizabeth Liddy, TextWise, Centre for Science and Technology, Syracuse University, Syracuse NY 13244-4100,
USA.

Mike Weiner is Chairman and Co-founder of TextWise, and also heads up Manning and Napier Information Services, a service bureau provider using the TextWise DR-LINK technology to help corporations and government agencies better manage their information assets. After a ten year career at Xerox, he founded Microlytics, a Xerox spin-off company which took technology from the Xerox Palo Alto Research Centre and created partnerships, alliances and licenses with over 100 leading software and hardware companies. Microlytics later merged with Selectronics, a publicly traded consumer electronics company, where Mike was CEO and Chairman.

Elizabeth D. Liddy is Associate Professor in the School of Information Studies at Syracuse University, where she also received her Ph.D. in Information Science. Her research focuses on the use of linguistic theory for information analysis. Dr Liddy has had more than 60 research papers published and has consulted for organizations improving their access to information. She is President and co-founder of TextWise, Inc.

Advances in machine intelligence, understanding, processing and analysis of information in text files, are all going to have a profound impact on business over the next five years. Information access will change dramatically, from information overload requiring intense effort to locate, read and analyze, to the equivalent of information radar that can spot relevant information, well beyond the current line of sight. Emerging retrieval systems will identify relevant documents, worded in a variety of different ways - and even identify trend lines, anomalies and new clusters of emerging and converging technologies.

Imagine a corporate intelligence information centre, just a few years hence, at a progressive high technology company located anywhere in the world. Traffic from disparate sources, newswires, internal electronic mail and research reports, Internet user groups, patent filings, corporate securities filings of companies in all the major world markets, traffic on the Internet at hundreds of sites, and a plethora of additional information is acquired, processed - more on this in a moment - indexed, analysed, and matched to the unique profiles and standing queries of thousands of corporate citizens, all in real time.

In the marketing department of a small division, a standing query has been loaded on the system to watch for:

"Any instance of anyone making, or proposing to make or distribute, chemicals for the treatment and preservation of steel bridges and steel civil infrastructure. This includes new product announcements, proposed new entrants, acquisitions or divestitures, and companies, individuals or Institutions announcing or implying new technologies or patents, affecting steel preservation, anywhere in the world."

This query is checked daily against the world wide flow of new information from all sources accessed. Whenever an article matches the profile, an e-mail is sent to the department information administrator,

containing the headline, source, date, lead paragraph and most relevant paragraph of the story; also included in this flash report are data on the companies and people mentioned in the report. The document itself can be ordered electronically and sent via e-mail, laser printer, fax or hard copy, within minutes.

In the public relations office at corporate headquarters, a standing query watches for:

"Comments by industry analysts in major publications [a term previously defined by the requester] commenting on the outlook for our company's [name of company and division would be entered] long term bond."

"Also of interest are comments on the outlook for the long term bonds of __, __, and __: [the company's main competition]."

In the legal department, the system is on the lookout for:

"Pending or proposed legislation or regulations that might impact, positively or negatively, industries involved in chemical processing and distribution within the United States, Mexico, Canada and Brazil. Included are environmental and occupational regulations, laws, ordinances or bans, including those affecting the shipment of chemicals by rail, truck, mail, air or courier. Not of interest are regulations or regulations affecting Europe, Asia or the Middle East. Report on this to me daily."

In the R&D department a researcher has an idea for a new method of waterproofing leather. He describes the invention and the applications envisioned, and runs a prior art search on world patents, doctoral dissertations, trade journals, product announcements, scientific and technical papers, and other sources, and receives back a stack ranked list of internal and external sources, organised at the push of a button by time, source, discipline, content, etc.

As patent searchers well know, searching prior art is not easy today. There are literally dozens - if not hundreds - of potential sources that running a robust search could involve. No one really can afford this, neither the patent offices, nor certainly inventors. But the introduction of a natural language search capability of the type suggested, working across a large collection of content sources, could fundamentally change patent searching as we know it. To seek out references which use different vocabulary, and to detect similar concepts, would reduce the number of instances of improper patents and make valid patents stronger.

Back in our hypothetical public affairs department, a notice has been sent alerting management to a sudden increase in the mention of the product of one of the divisions in a special interest group on the Internet. Normally, mention of this product across the net occurs approximately twice a week, but today there have been over sixty mentions. This could have positive or negative implications, reminiscent of the Intel Pentium flail of 1994. A PR person logs on, views the thread where the eruption has been detected, and is pleased and relieved to see a recent magazine review generating a positive dialogue about the company's new air filtering technology. The PR person sends a note to the product manager suggesting she log on and thank the group members for their kind remarks.

The ability to enjoy this type of capability without having to wade through numerous non-relevant returns is within months, not years, of realisation and dissemination. We are about to experience an improvement in intelligent text processing that will propel us from the equivalent of the open cockpit biplane to the supersonic stealth fighter.

The enabling technologies which will deliver these breakthroughs have come from a combination of private and academic institutions, many working for government. The Jargon for these new capabilities includes terms such as document detection, extraction, visualisation and routing. The problem has been a difficult one. Language is complex, and contains many ambiguities. Computers have had a difficult time performing the dis-ambiguation and understanding tasks that humans take for granted.

But years of investigation by psycholinguists into how we process and understand information, and years of work developing parsers and natural language grammars as computers have become more and more powerful, have all helped to move the technology along.

As with many significant technological advances, many of the underlying developments and breakthroughs have their genesis in US intelligence and defence programmes, funding academic and industry R&D. Information acquisition and analysis has been a growing strategic mission of the US government. We are about to experience an improvement in intelligent text processing that will propel us from the equivalent of the open cockpit biplane to the supersonic stealth fighter.

The authors have been involved with a US government initiative called *Tipster*, backed by the Advanced Research Project Agency (ARPA). The authors are commercialising the technology, known as DR-LINK (Document Retrieval through Linguistic Knowledge), through their company, TextWise. DR-LINK was created to help automate the process which research librarians use to convert information needs, as stated by users, into information search processes. (Prior to earning her Ph.D. in information science, Dr. Liddy worked as a research librarian).

Yet, the flow of information from disparate sources around the globe makes the task of getting the right information, regardless of subject, to the right analyst or decision maker, in real-time, a virtual mission impossible. Hence, this pooled effort, led by the Advanced Projects Research Agency (ARPA, formerly DARPA), spearheaded the first major organised scientific assault on the information overload problem.

It was an effort of impressive proportion, which is having a profound effect on technology for information access. In *Tipster* Phase One the US intelligence agencies pooled their efforts and formalised the quest for a solution for their information overload. In their case, national security mandated this drive; the socio-political, economic and strategic benefits of discovering an attempt to blow up the Holland Tunnel, or an equivalent target, the night before, as opposed to the morning after, can hardly be measured. Suffice it to say that the need has generated the investment.

The results of this technological investment have been significant. Substantial improvements in precision and recall performance were demonstrated. Yet, the puzzling phenomenon continues that two qualified searchers using two different retrieval systems can achieve comparable levels of precision and recall; and each will retrieve two collections of documents - from the same corpus - which have little overlap!

The net result of this observation is that two (or more) retrieval systems may be better than one; and our DR-LINK system actually contains several independent retrieval technologies that are blended into an amalgam of components that contribute to a single ranked list of documents, with a robust blend of relevant documents using different evidence sources.

PARADIGM SHIFT

This once esoteric and now often used term describes a fundamental shift in the baseline view, and in the case of the business information industry, such a shift is imminent. The technologies that are in development and in use in the intelligence community are radically different from those that are provided to business by the commercial online world of business information vendors. In the commercial world, online Boolean ('and/or') and statistical/probabilistic retrieval systems (such as PLS) are the norm. Customers are generally required to pre-specify the databases where they believe the information they are seeking resides, then give as detailed a query as possible, and are given back a set of documents which they sort manually for relevance.

Because these systems match on words and their expanded forms (plurals and synonyms), they do not differentiate on the multiple levels on which we, as humans, differentiate text.

Scientists in the field of psycholinguistics have determined that humans process text on at least six known levels (Figure 1):

morphological (the word bits, elephant vs. elephants, legal vs. illegal, etc.); **lexical** (the word forms themselves); **semantic** (various meanings associated with the word or phrase); **syntactic** (the ordering, C/S invades Panama, vs. Panama invades the US, there is a difference); **discourse** (all the material we derive above the word and sentence level, such as the referral further on in an article to information provided at the beginning; and on the difference between definitive information related to the past, and predictions and expectations in the future); **pragmatic** (that which we know, above the text: if you say third world, or NATO countries, I know what you mean; this is world knowledge, expected from others with whom you are communicating).

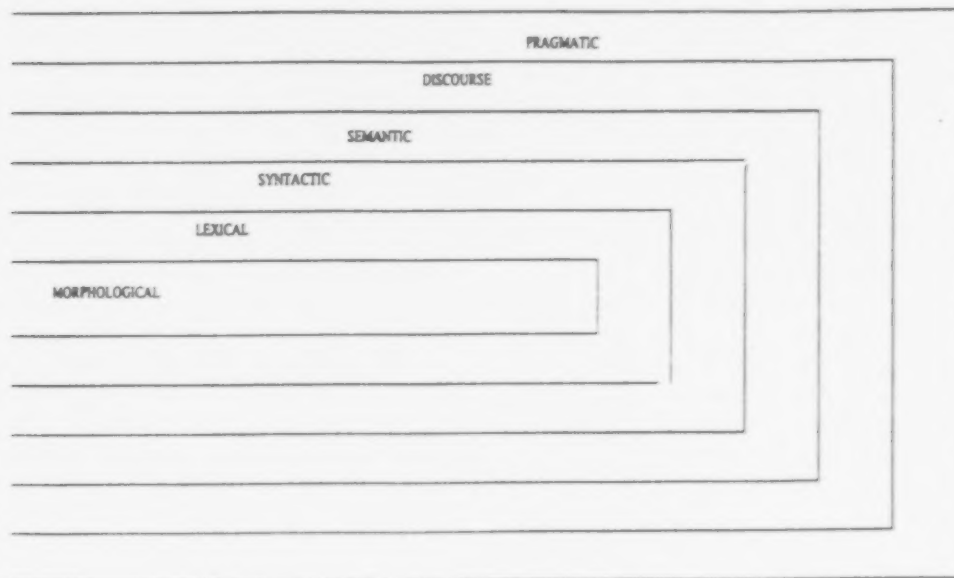


Figure 1: Synchronic model of language

One way to view the emerging class of intelligent text processing systems is that they attempt to process on these incremental layers. The more impressive of this new genre will attempt, when loading or indexing a document for future access, to do a form of intelligent note taking and annotation. Not only are companies and organisations noted, but known information about them is expanded and stored.

For example, if we encounter 'the White House' in a document about the US government or the President, we know that, in this context, we are describing an agency of United States government, located physically in Washington, D.C. in the United States of America, within North America, and we know that the US is also a NATO country. We also know this agency of government is not a European Union country, nor a third world country, and not located in South America.

The more sophisticated systems store this type of data bit in a relational database that allows you to ask what the US government is saying about a topic, and to pick up automatically statements issued by the White House, the Federal Reserve Board, and other agencies of the US government, without having to list them all.

If what you want are future, evaluative comments as to the outcome of a specific pending event, this is a different request from the historical outcomes of past events. Our studies of users and their queries, in intelligence and in investment work, frequently include this subtlety as a key portion of the user's stated information need. In fact, in the queries provided in the Tipster program, 85% included proper nouns.

Today's commercial online systems do not attempt to deal with this subtlety automatically, and therefore the effective use of these online systems requires trained search intermediaries whose function is to talk to these systems in their complex languages, and to also serve as filters to find for the information requester that which they need, as opposed to all that a commercial system will return.

This promises to allow us to search a large universe of information without having to pre-specify the database. Our experience indicates that this can produce many 'gold-nuggets' of information that might otherwise not be found.

And, a system which automatically interprets and expands on a user's natural language query can expand the access to information by members of an organisation, without the help of a trained search intermediary. Ideally, this frees up the trained searcher to focus on the most difficult and strategically vital searches, and to increase the overall usage of internal and external information within the organisation.

Unfortunately, we also see short-sighted efforts to eliminate the experienced search intermediary from some organisations, a loss whose long term implications can be severe, but difficult to measure.

PATENT SEARCHING

In the area of searching for prior art literature in the patent area, this type of linguistically informed search capability could change fundamentally the manner in which patents are prosecuted. If the patent office had access to such a powerful system, one able to overcome some of the complexities and obfuscations of language, what patent application could afford to be submitted without a comparable, comprehensive search?

If the search tool were powerful enough, one might be able to search not only newspapers, journals, magazines, abstracts and bibliographies, but also entire reference books, specialised encyclopedias, and even the plethora of doctoral dissertations that are currently searchable only in abstract form. These all represent significant amounts of prior art that we know may exist in the actual recorded, published universe, but which it has been impractical to search up until now.

At TextWise we are taking these opportunities seriously, and plans are under way to intersect them in the near term. On the horizon is the ability to enter a query in one language and access information in multiple foreign languages. Today, such a broad search is often impractical, unless litigation for infringement is under way. Again, this type of advanced research and development is enabled with R&D dollars from various government sources, but the implications for competitive intelligence, and particularly for patent retrieval, are substantial.

COMPETITIVE INTELLIGENCE

Today we have access to hundreds of local newspapers from cities, towns and hamlets across the world. However, if we are asked for every instance of mention of a competitor's company or product, we could

easily use up enormous resources trying to process all the information which would be returned, much of it not relevant to our interests. But if a standing query could alert us to a competitor's new or rumoured new products, and would also eliminate redundant information, we might find such a capability of immense utility in keeping track of the competitive terrain.

A system which can allow a decision maker to pose a natural language question such as those shown earlier, and get back reasonably relevant information, interpreted, expanded, and filtered to match what was wished for - as opposed to that which merely matched lexically - will open up a new level of service for industry, one that we believe industry very much wants.

All indications are that many of these breakthroughs in capability will come from new ventures with aggressive visions of these new capabilities, and not from the established and entrenched dominant information players of today.

INTELLIGENCE TRADECRAFT AND VISUALISATION

We expect also to see emerging new systems incorporating automation of proven intelligence practices, such as focus on 'POL (People, Organisation and Location); the relationships between the people in an organisation, the organisation and related organisations, and location.

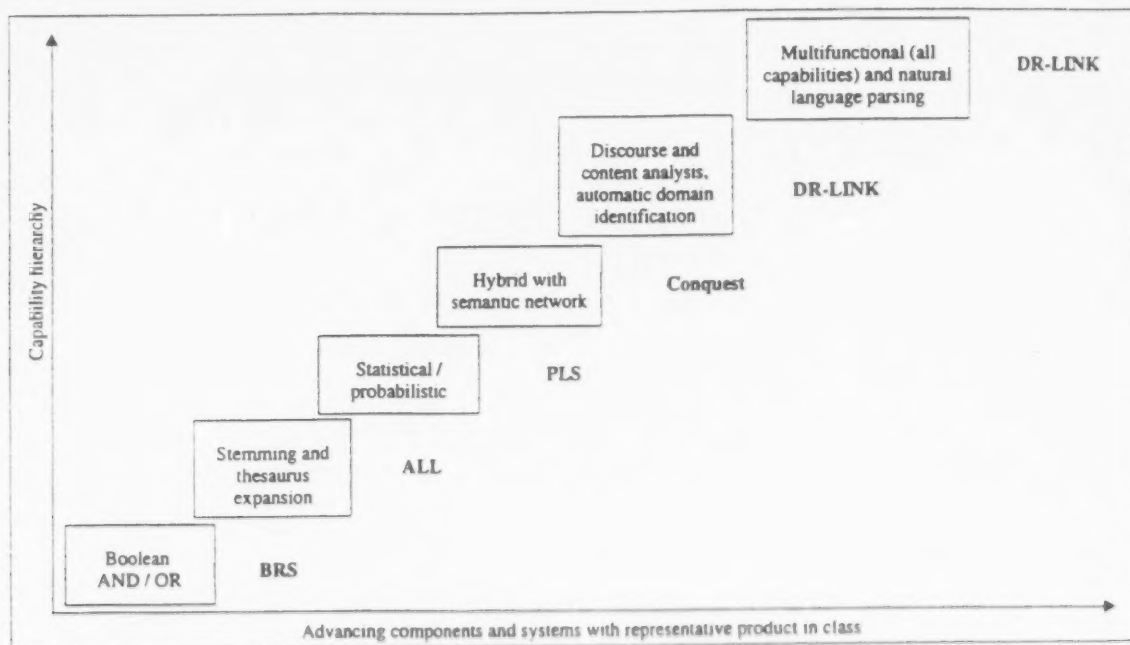


Figure 2: Progression of technical features over time

At the Xerox Palo Alto Research Center (PARC) in 1984, a project called Notecards was developed. In addition to being the conceptual predecessor to Apple's Hypertext and the World Wide Web, it was also the predecessor to using interactive machines to identify and visualise non-obvious linkages and sociopolitical influences within a given organisation. The project in 1984 analysed the Soviet Politburo. Dossiers on the members of the Politburo were entered into the system, including their place of birth, schooling, military service, political party affiliations, etc. Notecards then allowed different information to be viewed in a variety of ways.

We looked at one Politburo member's childhood home town, and were able to see that this member had grown up as a young boy in the same town as another Politburo member, and both had then served in the Russian military together. By viewing this town on a map, we noted that it was not far from the

home town of yet another Politburo member, and their careers had been linked in several places as they made their way up the organisation. Notecards was the first of a genre of visualisation tools that allowed viewing of not just the information, but how it interacted between people, organisation and location.

Technology has come a long way since Notecards. Today's visualisation tools allow multiple views of complex textual documents in 'docuspace.' A government visualiser called *Galaxies* has been used by intelligence analysts to identify unsuspected connections between suppliers of nuclear technology and Iran. Over 500 documents were analysed by viewing how they 'clustered' on screen, and queries were asked including 'China', 'Eastern Europe', 'Western Europe' and 'South America.' Where China had been the suspected supplier of nuclear technologies to Iran, this analysis identified Western European and South American suppliers that were not anticipated.

The human brain has an incredible ability to process massive amounts of information when viewing photos and visual scenes. A typical 35 mm print contains millions of bits of information, and we process these visually in nanoseconds, whereas, it takes hours - if not days - to read a few million bytes of text. The notion of using visualisation is to convert the textual patterns and information bits in documents into a form of 'Docuspace' to enable visualisation has been demonstrated in intelligence practice. It works, in amazing ways. Look for this capability to migrate to the competitive intelligence and business intelligence function in corporations, and to make a noteworthy difference to investment analysis.

With intelligent text processing we can expand the visualisation to analyze things above the word level. In our DR-LINK system at TextWise we have a technology which converts the words in a document into a higher level concept group, allowing us to identify that a specific document consists of 44% economics, 17% biotechnology and 12% business law lexical coverage; this concept clustering of a document that might contain 2,000 words allows a visualisation to be able to group, in a logical way, the content of thousands or even millions of documents.

In intelligence trade craft we can use this capability to guess what personnel at a specific research facility are really working on. In investment analysis we might use such a tool to analyze, both visually and statistically, changes in the makeup of content over time. Once patterns of relationships and events can be visualised, tracking systems can be put in place to alert us automatically to changes in state.

These enabling technologies have numerous implications for our ability to keep track of the rapidly growing amounts of information that are available to us, and are also available to our competition.

In the personal computer industry, fortunes have been made and lost on being on or off cycle on critical trends. Noticing the shifts and being up to speed on operating systems helped Bill Gates take Microsoft from a sleepy little company producing a Basic software compiler to the colossus it has become. Watching trends in multimedia have helped companies stay on top of developments in CD-ROM growth, graphics boards, sound cards and other peripherals.

AUTOMATED TECHNOLOGY TREND SPOTTING

We like to think that our current technology, had we had it in place in the 1950s and 60s, could have predicted first the emergence, and then the rapid growth, of the semiconductor industry. Never before had there been a culmination of technologies as disparate as transistors, photolithography, silicon crystal growing, wave soldering machines, and other developments, impacting the explosive integrated circuits industry.

Having the clues is not enough. Once you determine what you believe the expected outcome will be, you need a means to monitor your assumptions. The same technology that helped visualise the trend, and for the analyst to make the prediction, can now be used to monitor the trend, as events unfold.

If our belief proves valid, there are potentially many other applications for such an analyser and predictor. We need this, since it is not practical to keep track of the growing deluge of information sources with filters that work only on key words and do not distinguish between the future and the past, nor the difference between an electrical circuit and a circuit court.

INTELLIGENT AGENTS

This has become one of the more exciting developments anticipated for navigating the information highway. There are many ways to enhance the performance of such an agent. On an information source as rich as the Internet, the agent must move past the key word or key word plus synonym type of search tool that is available today. The agent must be able to discern these additional levels of linguistic evidence.

Just how intelligent the intelligent agent proves to be, will dictate how widely used, and how useful, this emerging technology proves to be for the decision making process in business, industry and finance.

It is our prediction that the utility of these tools will be substantial, and for the most part, exceed the expectations of today's users and purveyors of business information.

BREAKING THE LANGUAGE BARRIER

As we look further out on the horizon, our intelligent agent will no longer be bound to searching only those languages which we speak. The task of actually translating all the documents in docuspace appears daunting, but the ability to first search in multiple languages, and only have to translate those that are relevant, would allow us to keep track of emerging new talent and creative start-up companies, and potential competitive threats, on a global basis.

NEVER BEFORE POSSIBLE

The combined advances indicated in this article are the result of creative new approaches in information retrieval, linguistics and knowledge based systems. However, without the significant advances, cost reduction and dissemination in machine processing capability, these developments would not have been possible. In our development laboratory we use a powerful Sun SPARC multiple processor server with 512 megabytes of RAM and 70 gigabytes of hard disks as a server, linked to an even more powerful IBM SP-2 Massive parallel Processor with 12 nodes, accessing over 600 gigabytes of information, and connected again to an even larger system with literally hundreds of powerful processors.

This type of information processing horse power was just not available to us only a few short years ago. The incredible technological developments across multiple fields of computer and software technology are allowing an attack on the problem of information overload and information analysis that will pay big dividends in the area of global strategic intelligence in exciting ways, and in the not very distant future.

This paper was first published in *The Journal of AGSI* July 1995 (ISSN 0965-4380).

ADVANCED SEARCHING: TRICKS OF THE TRADE

Peggy Zorn, Mary Emanoil, and Lucy Marshall
Parke-Davis Pharmaceutical Research Library
and Mary Panek
United Technologies Research Center

Parke-Davis Pharmaceutical Research Library, 2800 Plymouth Road,
Ann Arbor, MI 48105, USA
Tel.: +1 313 996 7202, Fax.: +1 313 996 7008, E-mail: zormm@aa.wl.com
United Technologies Research Center, Mail Stop 169-31, East Hartford, CT 06108, USA

INTRODUCTION

Searching the World Wide Web has become as easy as clicking on the Net Search button in Netscape. Type in a few keywords to describe what you're looking for, click on the Submit button and in a few seconds you have your answer. . .or do you?

Despite the rapid proliferation and development of Web search systems over the past months, little attention has been devoted to the advanced features professional searchers and librarians have become accustomed to in other online information resources.

Web novices and information professionals alike often overlook or miss detailed information such as:

- what a particular Web search system is searching
- how the data has been indexed
- how the search engine retrieves data
- what advanced techniques (proximity operators, nested queries, search set manipulation and combination, duplicate detection, etc.) are available

In simple Web systems such as Yahoo! and Aliweb, knowing advanced search features is not as necessary because the depth of indexing and the power of the search engine is not as great. However, with more sophisticated Web search systems, the need to narrow basic keyword retrieval due to large retrieval sets using types of features available in established commercial online services is becoming painfully obvious.

As the Web grows, and databases searched by the major Web search engines increase in size, the power to seek, index, and retrieve information must grow - along with the information professional's knowledge of how to effectively search the Web. Typical end-users may have no trouble browsing and locating information on uncomplicated topics, but constructing complex search queries using sophisticated Web search engines is another matter entirely. End-users may increasingly rely on information professionals for complex Web searching, much as they do for online commercial databases.

ADVANCED FEATURES IN WEB SEARCH SYSTEMS

Not all Web search systems, their search engines, and the databases they search are created equal. Only a handful really cover the Web exhaustively in terms of URLs covered and depth of indexing. Only a few provide the type of advanced searching features that are standard for librarians.

In the November/December 1995 *ONLINE*, Martin Courtois, William Baer, and Marcella Stark provide a comprehensive overview of many search systems on the Web and detail the results of performance testing on sample searches. Several times the authors state that, even in the search systems that offer advanced searching features, instructions for use were either difficult to locate or difficult to comprehend. The good news is that there is a trend toward more search features and simplified, advanced searching documentation. These enhancements promise to improve Web searching accuracy and efficiency.

Besides advanced searching features to refine retrieval, a database's size and depth of indexing also greatly affects the complexity of searching that can be done using a particular engine - and the resulting retrieval. Some engines support full-text searching of the Internet sites included in their databases. Others, such as Magellan and Open Text, have gone further by indexing on the field level (title, author, abstract, keyword, etc.). They have also added descriptors and other information not necessarily available on home pages to aid searching and retrieval. This type of field and descriptor searching and value-added indexing is a mainstay of online commercial databases, and likewise is of great use for narrowing broad topics and filtering retrieval in Web searches.

EVALUATING ADVANCED SEARCH FEATURES

The purpose of this article is to look closely at several Web search systems that provide advanced search features and search a comprehensive and authoritative database of Internet sites. Based on these two key requirements, we selected Alta Vista, InfoSeek, Lycos, and Open Text for evaluation and testing.

The search features we looked for included:

- complex Boolean (nested logic)
- duplicate detection
- keyword(s) in context
- limiting retrieval by field
- proximity and/or phrase searching
- relevancy ranking of results
- retrieval display options
- search set manipulation
- truncation (automatic or user-defined)

We also looked closely at depth of indexing and quality of documentation, as well as the size of the database the engine searched. Only search engines with a database of more than 200,000 Internet sites were considered. Initially, we considered several other Web search engines and databases, including Harvest, Magellan, NlightN, and Yahoo!. NlightN and Yahoo! both search significant databases, but their search engines are not powerful enough to handle our queries effectively. Harvest and Magellan, on the other hand, provide powerful search engines that include most of the features listed above. However, their databases are too small at this point to really label comprehensive.

Here is a detailed description of the four Web search systems we studied - their features, how to use them, and how the search engine performed on our sample searches. Each sample search requires more than a simple keyword search and tries to incorporate as many advanced features as possible. All three searches were run on all four search systems, varying the syntax based on features available on each. The following table shows a quick overview of the features of the four systems.

	Alta Vista (altavista.digital.com)	InfoSeek (www2.infoseek.com)	Lycos (www.lycos.com)	Open Text (www.opentext.com)
Approximate # of URLs	16 million	1 million	10.75 million	1 million
Documentation	Excellent; includes details and search examples for both Simple and Advanced queries	Lengthy, somewhat difficult to locate	Fair, not as thorough as it could be	Excellent, explains all features in detail
Duplicate Detection	No	Yes	No	Yes
Field Searching	Yes	No	No	Yes, choose field limits from pull-down menus
Indexing	Full-text	Full-text	URLs, also choice of other parts of pages and text	Full-text
Multiple Search Sets	No	No	No	No, but allows up to four Boolean or proximity operators per query
Nested Boolean	Yes	No	No	No
Proximity Searching	Yes, ensures both terms are within ten words (or phrases if terms are enclosed in quotes)	Yes	Yes, choose match type (loose, fair, etc.) from pull-down menu	Yes, choose operators from pull-down menu
Relevancy Ranking	Yes, can specify which terms to weight first in retrieval display	Yes	Yes	Yes
Truncation	Yes, can use an asterisk on word stems longer than three letters	No	Yes, automatic	Yes, automatic

SAMPLE SEARCHES

Search #1: Locate information on cancer research grant or funding opportunities. This search could include truncation to retrieve alternate word endings and more than one Boolean operator to combine synonyms and concepts. On traditional commercial online services, this search might look something like this:

(cancer or oncol) and research and (grant* or fund*)*

Search #2: Locate information on Warner-Lambert or its pharmaceutical research division, Parke-Davis. This search is tricky because of the embedded punctuation. This problem can be handled in several ways, based on the command structure of the search engine. A typical online search for this concept might look like this:

(warner adj lambert) or (warner-lambert) or (parke adj davis) or (parke-davis)

Search #3: Locate information on the XI International Conference on AIDS. This search has numeric content, and requires the use of synonyms and multiple Boolean operators. A search for this topic might be entered like this:

(xi or 11th or eleventh) and international and conference and aids*

Alta Vista

<http://altavista.digital.com>

A relative newcomer to the Web search service arena is Digital Equipment Corporation's Alta Vista, introduced to the general public in mid-December 1995. Alta Vista contains full-text indexing for approximately 16 million URLs in its database, and provides a powerful and flexible search engine for performing queries. Probably the most refreshing thing about Alta Vista from an information professional's perspective is the similarity the search commands and features have to traditional online commercial database services. Use of Boolean operators (AND, OR, NOT, NEAR), the ability to do nested Boolean queries, and the ability to use truncation when needed (but not have terms truncated automatically by the search engine) are examples of such familiar features. Documentation for the searching features available in Alta Vista is detailed and filled with plenty of examples.

Alta Vista offers two ways to search its database, Simple and Advanced. The Simple search does not support Boolean operators, but uses a combination of different syntax to indicate phrases, proximity, and required or prohibited words. Limiting searches by field (Title, URL, Host, etc.) and truncation is possible in both modes. Results from queries done in the Simple mode are returned in relevancy-ranked order.

The Advanced search uses the same syntax as the Simple search for defining words, phrases, wildcards, and punctuation. However, with the Advanced search, Boolean operators must be used to combine words and phrases, and parentheses are required to control nesting. The retrieval can optionally be ordered (by specifying a set of words or phrases to list first) and can also be limited by date.

Search #1: The Advanced search mode was used for all the sample searches. The search for information on cancer research grants was originally entered as:

(cancer or oncol) and research and (grant* or fund*)*

An error message appeared indicating too much retrieval for the truncated forms of "grant" and "fund," so the search was rephrased using the following strategy:

(cancer or oncol) and research and (grant or grants or fund or funds or funding)*

This search retrieved over 20,000 documents that matched at least some of the terms. Specifying "cancer" and "research" in the Relevancy Ranking Criteria area of the form improved the precision of the items retrieved at the top of the returned listing.

Search #2: The search for information on either Parke-Davis or Warner-Lambert used the internal punctuation conventions in Alta Vista. Punctuation characters in words or phrases act as separators, splitting the words into separate entities. In addition, capital letters entered in a search query are considered distinct from lower case. Usually, terms should be entered in lower case to indicate a case-insensitive match, unless only retrieval of the capitalized word is desired. The search strategy was:

"Parke Davis" or "Warner Lambert"

This search retrieved approximately 2,000 hits, with the Warner-Lambert home page listed within the top ten items.

Search #3: Searching for information regarding the XI International Conference on AIDS was entered as:

(xi or eleventh or 11th) and international and conference and aids

The terms "aids," "international," and "conference" were entered as criteria for relevancy ranking the results. Over 1,000 hits were retrieved and the home page for the conference was listed first.

InfoSeek

<http://www2.infoseek.com/>

InfoSeek was introduced in February 1995 by InfoSeek Corporation and was the first search system on the Web to charge a fee for searching. InfoSeek currently offers two ways to search for information. Users can access its collection of Web pages, newsgroups, FTP, and gopher sites and retrieve up to 100 hits for free.

InfoSeek membership plans are available for those who want to access InfoSeek's additional fee-based databases, including Usenet News; Cineman Movie, Book, and Music Reviews; Hoover's Company Profiles; a selection of wire services; CorpTech Directory of Technology Companies; and MDX Health Digest. New databases in a wide variety of areas, including business, finance, health, sports, and national news are added continuously. Membership plans vary from a flat \$0.20 per search to \$9.95 for a monthly subscription. The paid subscription also offers access to many full-text articles. Although the search features are the same for the free and fee-based databases, we tested only the free service.

InfoSeek currently includes indexing for over one million Internet sites. It uses a combination of Web "robots" or "crawlers," as well as user submissions to populate its database. Users can easily submit their sites to InfoSeek for inclusion in the database. They do not need to provide a list of keywords, but are asked to make sure their documents contain words that accurately describe contents because InfoSeek indexes the full text of all sites included in its database.

Query results are returned by software that uses statistical word frequency counts and other techniques to determine significance and relevancy. Although InfoSeek does not provide much further explanation of its rating scheme, it stresses that the rankings are based on more than word frequency counts. In addition, InfoSeek attempts to eliminate duplicate entries for URLs in its database and claims never to index a URL more than once. URLs with duplicate titles are allowed, but fewer duplicates are retrieved by an InfoSeek search than from other Web search systems.

InfoSeek provides a help page that explains how to perform searches against its databases. It contains thorough explanations and good examples on Simple searching, using advanced query operators and syntax in search strategies, and refining search techniques. There is also an extensive FAQ section, but it is somewhat buried.

InfoSeek has a number of advanced search features. Unlike many other search engines on the Web, InfoSeek does not support Boolean operators or wildcard operators. When using advanced features in InfoSeek, users need to apply special syntax such as:

- double quotation marks to find words that appear next to each other
- hyphens between words to locate those words very close to each other
- brackets around words to retrieve those words near each other in any order
- a plus sign in front of a word or phrase to return that specific word or phrase
- a minus sign in front of a word or phrase to ensure that word or phrase does not appear in any of the resulting documents

The search engine for InfoSeek is case-sensitive. Capitalizing words forces a search exclusively for capitalized instances of that word.

Search #1: InfoSeek does not support Boolean operators or wildcard characters, so Search #1 for information on cancer research grants could not be entered as it might be on a traditional online service. In our first attempt at this search, the following terms separated by hyphens were entered:

cancer-research-grants

This search strategy only returned three documents, of which only one was relevant. Hyphens require that the terms appear very close to each other, but we overlooked the fact that the hyphens also limit the order of these terms. Therefore, the next search strategy included the same search terms but with brackets around the three words to retrieve those words near each other in any order.

The bracketed search retrieved 100 documents - the maximum number allowed in a free search - and the results were more on target.

Search #2: Not knowing how InfoSeek handled embedded hyphens, we tried the search for information on Warner-Lambert and Parke-Davis companies in two ways. For the first search, the query was entered as:

"Warner-Lambert", "Parke-Davis"

The quotation marks forced retrieval of words that appear next to each other. This search retrieved 100 documents. Surprisingly, the Warner-Lambert home page was not the highest-ranked document. It was ranked 16th, behind some documents that merely mentioned Parke-Davis as a previous employer. The sixth-ranked document did not show either of the two companies on its first page, nor did it contain an obvious link to either one. To determine the importance of the hyphen within the company names, we did a second search without the hyphens:

"Warner Lambert", "Parke Davis"

This search retrieved the same set of 100 documents. Apparently, InfoSeek ignores embedded punctuation, at least hyphenated punctuation.

Search #3: To locate information on the announcement for the XI International Conference on AIDS to be held in 1996, we formulated a search using a few keywords from the full phrase. We put hyphens between the terms to require close proximity and force the order of the terms.

11th-conference-AIDS

One document was returned using this query and it contained a brief announcement about the conference. However, 11th could be presented as "eleventh," "11th," or "XIth." Since Boolean operators are not available, we had to submit three separate searches to pick up the variant spellings of the term. For the second search, the following query with the Roman numeral was used:

XI-conference-AIDS

The strategy found 31 documents and the first three were direct links to the actual home page for the 1996 conference. No documents were retrieved with a strategy that spelled out the term (Eleventh-conference-AIDS).

Lycos

<http://www.lycos.com>

As indicated by its Latin name for WolfSpider (Lycosidae), Lycos is a spider-oriented search engine. It was developed by the School of Computer Science at Carnegie Mellon University, and is now owned by Lycos, Inc., a joint venture of CMG@Ventures and Carnegie Mellon University.

Lycos systematically rebuilds its database daily and checks for active links. The search engine provides weighted retrieval from the database and returns a user's query with hits sorted by relevance ranking. Rankings are determined by the location of the word in the document. A keyword located in the title or header receives a higher ranking than a keyword located in the document text.

The search results include the match score, number of links, document title, headings, sample abstract, the URL, and length of document. Lycos boasts an index of over ten million pages throughout the world - indexing over 91 percent of the content of the Web. With the Lycos catalog growing by over 300,000 pages per week, it claims it will soon catalog 99 percent of the WWW.

Recently, Lycos has created a search form that offers easy access to search and display options. The Search Option button lets you select AND and OR as well as the level of closeness or looseness of a match. It has modified advanced Boolean searching with the ability to combine two out of three search terms. This match is not as powerful as true Boolean searching, but does allow multiple concepts and synonyms to be combined. Using the Display Option button, the user selects how many hits are desired and the level of display detail for the results (standard, summary, or detailed). The system automatically truncates search terms.

Documentation for searching and displaying is available, but is not extremely detailed. Lycos describes how to do a variety of searches, but does not explain how they are completed. For example, there is no real explanation of the difference between a "close" or a "loose" match other than the difference in number of hits returned. Repeated E-mail messages sent to Lycos asking for further explanation were unanswered.

Search #1: The first search was entered (with search options set to match three terms) as:

cancer oncol research grant fund

Presumably, oncology will not appear if cancer is present, and funding will not be a term if grant is present. The retrieval was more accurate than expected (146 documents). The first few results included several highly relevant sites including the NIH Grants Database and the FAA Research Grants Program. Several of the documents ranked lower in the retrieval list contained the terms "research," "grant," and "fund," but not "cancer" or "oncology." Because the ability to do true Boolean searching with nested queries is not available in Lycos, several searches using different synonyms and combinations would need to be performed to gather comprehensive results.

Search #2: In the search for information on Warner-Lambert or Parke-Davis, Lycos ignored the hyphenation both in the search strategy and retrieval. This was beneficial because, although a search on Parke-Davis typically retrieved Parke Davis, it also retrieved Parke-Davis.

Constructing the search strategy for both phrases caused the same problems as the first search, where the search options had to be set to match two terms. This search engine restraint causes terms to be ANDed together, which often leads to results like we found in one of the hits - a music page discussing Bobby Parker and James "Thunderbird" Davis. This example also demonstrates the automatic truncation feature in Lycos, which is good for searches that require exhaustive retrieval, but may result in false hits.

Search #3: The third search revealed an interesting problem with the Lycos search engine. It currently does not search for numbers. This search had the potential to provide an example where the Lycos Search Options work as explained and expected. In looking for information on the XI International Conference on AIDS, it was easy to specifically select a match on any two terms and ask for:

11 11th XI eleventh aids

It is highly unlikely that a document will contain both "eleventh" and "11th." In practice, though, only the terms "xi," "eleventh," and "aids" were searched for. This search in Lycos produced two hits, both pointers to the conference home page.

Open Text

<http://www.opentext.com>

Developed by Open Text Corporation of Waterloo, Ontario in partnership with UUNET Canada, Open Text is a full-text index of over one million Web sites as well as FTP sites and gopher servers. Open Text uses a crawler to scan Internet sites and then uses Open Text 5, its proprietary indexing software, to index every word of each Internet site. The results are then stored in the Open Text database. New sites are added daily and old sites regularly revisited to update bad or changed links.

The power behind the Open Text is the index itself. According to the documentation, Open Text 5 can index more than 40 types of files. It indexes each word of the page so no stop words are ignored in search statements, identifies fields where words appear, and is multilingual. In addition, several technologies make Open Text 5 useful for organizations that require a powerful text-retrieval system. Using the product to index the Internet allows Open Text to demonstrate its features as well as provide a useful service to the Internet community.

Open Text offers three types of searching, Simple, Weighted, and Power Searches. The Simple Search is a basic keyword search with options to search for the exact phrase entered, all of the words (implied AND), or any of the words (implied OR).

The Power Search is more sophisticated and allows the user to select where the word(s) are searched:

- Anywhere (everywhere on Internet)
- Summary (combination of title page, first-heading, and text Open Text has deemed important)
- Title (indicated by the Web page author)
- First-Heading (indicated by page author)
- URL
- Hyperlink (outbound links from Web pages)

Basic Boolean operators can be selected, as well as proximity operators. The Weighted Search allows the user to assign importance to a word or phrase. As with the Power Search, it is possible to select where in the Web page the search will be conducted. A number is then entered in the weight box to assign importance, with higher numbers designating greater importance.

Of the three types of searches, Power Search is the most advanced. It provides the searcher with an easy-to-use interface, Boolean and proximity operators, and field searching. As you will see in specific searches here, the lack of nesting capabilities does present problems. The online documentation discusses the index's limitations with parsing queries. Evidently, it does not allow for parentheses in the user interface, so terms left at the end of a search statement are often dangling. Open Text suggested conducting two separate searches to remedy this problem, but this approach did not adequately identify the types of information being sought. According to Open Text, search set manipulation and the ability to save searches are features that will be added in the near future.

Results are displayed in a relevancy ranking for both the Simple and Power Search. The Weighted Search allows the user to display results by occurrence count (the product of the number of occurrences of the search terms and the weight of the terms), or the presence/absence of search terms (calculated by the weights given for the terms without regard to the number of search term matches).

Open Text supposedly has some ability to eliminate duplicates, but flaws have been reported. In our test searches we did not find duplicate listings. Each matched page offers the option to visit the page, see matches on the page, or find similar pages. The excellent online documentation explains all of these features in detail. The instructions for using the various search modes and the Open Text FAQ provide ample information for most types of searches. Open Text also offers a feedback button for questions that have not been addressed in the documentation.

Search #1: The search for information on cancer research grants was conducted using Power Search and Weighted Search. Initially, terms were entered in the Power Search as:

cancer or oncol and research and grant or fund

Truncation is automatic unless a space is entered following the term which stops truncation. Searches in First-Heading, Summary, and Title each produced several hits but relevancy was limited. After analyzing the search and the Open Text documentation, it was apparent that using "or fund" as the last term had the effect of combining the previous terms and then ORing them with "fund."

The search was conducted again without "fund" as a search term. There were fewer than ten hits in each of the three search areas, but only the Summary search produced links of identifiable relevancy. Eliminating "oncol" did not alter the results. Still, it seemed that there should be sites of greater relevancy so a search was attempted using a proximity operator. We entered the search as:

cancer near research near grant

We selected the same areas, as well as "Anywhere on the Web page" and found surprising results. Neither First-Heading or Title retrieved any sites, Summary retrieved only two, and Anywhere found 73 matches. The results found in Summary were highly relevant and were also the top two hits of the 73 in the Anywhere search. It is interesting to note that the sites located in Summary contained the search terms in the title of the page, but why the pages were not retrieved in a Title search is unclear.

The same type of search was conducted using the Weighted Search. Since only four terms can be entered, the following terms were searched, and the weights are in parentheses - cancer (20), research (10), grant (20), fund (5). Results were ranked by occurrence. A search of Anywhere on the page produced 786 hits while searches of Summary, Title, and First-Heading produced fewer than five hits each. None gave exceptionally relevant results.

Search #2: A search for Warner-Lambert or Parke-Davis was conducted using Simple Search and Power Search. Hyphenating the terms did not produce any hits, so the terms were entered without a hyphen. The Simple Search was entered as:

warner lambert parke davis (any of these words)

Over 30,000 pages were found, but the Warner-Lambert home page was not among the first ten. Trying the same search but asking for all the terms produced 35 pages, none of which were the Warner-Lambert page. Using Power Search the terms were entered as:

warner followed by lambert or parke followed by davis

This search found 89 matches. Anywhere on the Web page, many of which were relevant to Warner-Lambert Parke-Davis, but again the Warner-Lambert page was not among the pages found. Limiting to Summary, Title, or First-Heading retrieved fewer than seven hits each - again without find the Warner-Lambert page. Analyzing the search it was apparent that the last term, "davis" was dangling. To remedy this, the search was reconstructed as:

warner followed by lambert or parke davis (where parke davis is a phrase)

Performing this search in the Title field retrieved one hit, the Warner-Lambert home page.

Search #3: The search for the XI International Conference on AIDS was also conducted using Simple Search and Power Search. Wording for the conference number was entered as either 11th or Eleven or XI, along with various combinations of the other terms. Results included sites that linked to the conference home page as well as to the conference home page itself.

MAKING CHOICES FOR COMPLEX WEB SEARCHING

After looking at the features of the four search systems covered above and reviewing their performance on the sample searches, the conclusion was reached that no single Web search system is really "the

best." None of the four systems can claim to include *all* of the Internet in their databases. And, because the indexing and value-added evaluation process for including sites is so different, no system can claim to have it all at this point.

Alta Vista and Lycos are probably the most comprehensive in terms of number of URLs included. But the value-added indexing and selection process for URLs included in Alta Vista, InfoSeek, and Open Text lead us to rate them more highly in terms of relevancy and accuracy of retrieval. Open Text scores very well for including the most advanced search features, and for providing a good user interface and excellent documentation for the more complex functions.

Some search features expected in most commercial online services, such as search-set manipulation and duplicate detection, are missing from most Web search systems at this point. Full-text indexing of Web pages, including indexing of embedded links, create huge databases of Web sites that have hundreds of duplicate entries, explaining the large number of duplicate hits often retrieved with Lycos searches.

As any good searcher knows, complex queries requiring multiple Boolean operators can be entered in one statement in most commercial online searching systems. However, a searcher normally prefers to split concepts and operators into multiple search statements. Only one search window for entering terms is available in all but the Open Text system, making separation of concepts and combination of terms tricky. Even in Open Text, where it appears that nested logic is available, Search #1 proved that using an OR statement at the end of query skews retrieval.

PROFESSIONAL SEARCHING ON THE WEB

How should the professional searcher approach complex Internet searches? The best approach, given the current tools, is probably one that includes all of the four search systems covered here. Just as a searcher searching online commercial databases often runs searches on more than one database or search service to ensure comprehensiveness and accuracy in retrieval, the same is true for Web searching. Information professionals need to become familiar with the advanced features of these major Internet search services, and make an effort to have input into their future development.

Information professionals should closely watch developments in the Internet searching arena over the next few months. The licensing of the Lycos search engine by Microsoft, of WebCrawler by America Online, and the increasing popularity of fee-based Internet searching services such as InfoSeek and NlightN point to a future for Web search systems that may more closely resemble the current online commercial searching environment. Commercialization of some of the major Web search systems may ultimately lead to more powerful, accurate, and efficient searching for professionals and end-users alike.

END USERS TURN TO SEARCH EXPERTS

In many ways, current Web searching compares to the beginning of end-user online and CD-ROM database searching. Evaluations of end-user searching indicate that end-users like to do their own searching, but still depend on trained information professionals for complex or comprehensive searches. A recent survey by J. Fisher and Susanne Bjørner gauged the effect of wide-spread end-user searching on mediated searching. It showed that although a large percentage (85 percent) of corporate sites provide end-user searching, most (91 percent) of these corporations' librarians said that: "end-users continue to rely on professionals for all but the most basic research" [1].

Even with a variety of searching tools and extensive training, evaluator P.T. Bysouth concludes that scientific end-users have relatively simple queries [2]. Others agree with these conclusions [3,4,5].

Also, many studies show that libraries are often seen as more crucial with the advent of end-user searching - rather than of diminished importance. "The work in the information department has increased rather than decreased in spite of the introduction of end-user searching" [6].

It is clear that end-users will continue to call on information professionals, whether searching takes place on a traditional online service or on the Web. As more end-users search the Web, it is up to information professionals to be familiar with the best advanced Web searching tools, the most effective methods for searching, and the most accurate query structure to handle the more difficult searches and be comfortable teaching others. Becoming expert in the more robust features of Web search engines, and participating in their future development (building on the lessons learned in online searching) is an important role for information professionals, now and in the future.

REFERENCES

- [1] Fisher, J. and Susanne Bjørner. "Enabling Online End-User Searching: An Expanding Role for Librarians." *Special Libraries* (Fall 1994): pp. 281-291.
- [2] Bysouth, P.T. "Evaluating the Use of Several Approaches to Online Literature Retrieval by Research Scientists." *End-User Searching: The Effective Gateway to Published Information*, Association for Information Management (1990).
- [3] Crea, Kathleen, Jan Glover and Majlen Helenius. "The Impact of In-House and End-User Databases on Mediated Searching." *ONLINE* 16, No. 4 (July 1992): pp. 49-53.
- [4] Martin, H. and D. Nicholas. "End-Users Coming of Age? Six Years of End-User Searching at The Guardian." *Online & CDROM Review* 17, No. 2 (1993): pp. 83-89.
- [5] Brody, Roberta. "End-Users in 1993: After a Decade." *ONLINE* 17, No. 3 (May 1993): pp. 66-69.
- [6] Butcher, H. "The BDO Binder Hamlyn End-User Project." *End-User Searching: The Effective Gateway to Published Information*. Association for Information Management (1990).

SOLVING SEARCHING MYSTERIES AND FALSE DROPS

Most searchers like to try a few quick searches to test a new search engine. Who can resist typing in a few terms into a blank box and seeing what comes up? Often, sifting through a pile of less relevant material, you find some truly interesting results. Equally as often, something appears that makes you wonder where it came from.

Search engines on the Web have some defaults and basic features that are different, and thus not intuitive, for professional searchers. These settings are often the culprits that cause those unusual false drops.

For example, the default for many Web engines is to OR terms together, then provide results based on relevancy. This combination produces a retrieval that has all terms present in the first few hits, and then fewer terms as you move through your hit list. This explains why, even though your terms were ORed together, the first hits are exactly what you are looking for, and the last few do not even contain all of your search terms. Unless you specifically ask to AND terms together, do not trust your search retrieval number to accurately portray the number of hits from your search strategy.

Another typical default is automatic truncation on each term. So if your search is for "federal register," you will also retrieve documents with the terms "registering," "registered," and "federalist" in them.

Another way of explaining false drops is by determining exactly what the search engine is searching. Usually, the default is the URL, but sometimes a search engine retrieves documents where your search terms appear anywhere in the document. An address might include your search term, but the actual document may not show your term. Pay close attention to any Web engine documentation to clarify just how and what it searches. Those searching mysteries can be solved!

ZORN: This paper first appeared in *Online* May/June 1996, p. 15-28 (ISSN 0146-5422) and has been reproduced here with the kind permission of the original publishers: Online Inc., 462 Danbury Road, Wilton CT 06897-2126, USA.

WHAT DO PEOPLE WANT FROM INFORMATION RETRIEVAL?

W. Bruce Croft

Centre for Intelligent Information Retrieval, Computer Science Department,
University of Massachusetts, Amherst, MA 01003-4610, USA
E-mail: croft@cs.umass.edu

INTRODUCTION

With the enormous increase in recent years in the number of text databases available on-line, and the consequent need for better techniques to access this information, there has been a strong resurgence of interest in the research done in the area of information retrieval (IR). For many years, IR research was done by a small community that had little impact on industry. Most applications of text retrieval focused on bibliographic databases, and the large information services such as DIALOG or WESTLAW were based on standard Boolean logic approaches to text matching and paid little attention to the results of research on topics such as retrieval models, query processing, term weighting and relevance feedback.

Today, however, the situation is considerably different. Retrieval techniques based on IR research have found their way into major information services (for example, West Publishing's WIN system, Individual's clipping service) and the World Wide Web (for example, InfoSeek and Lycos). Many of the features once considered too esoteric for the typical user, such as "natural language" queries, ranked retrieval results, term weighting, "query-by-example" and query formulation assistance, have become common and, indeed, necessary in most IR products (PLS, Verity and Fulcrum).

Given the speed with which industry has adopted the results of IR research from the 1970s and 1980s, the IR community is faced with identifying major new directions. The emergence of new applications such as "digital libraries" is both an opportunity and a challenge. These applications provide unique opportunities as testbeds for evaluating and stimulating research, but the challenge for IR researchers is to define and pursue research programs that maintain their relevance in a rapidly changing environment. One problem is that the priorities that IR researchers place on research issues are not necessarily the same as those of companies and government agencies that use and sell IR systems. Understanding those priorities and the operational experience behind them will be part of the process of deciding which issues are of fundamental importance and which are more transient.

In this paper, I summarize the experience of the National Science Foundation (NSF) Center for Intelligent Information Retrieval (CIIR) in the area of industrial and government research priorities. The Center has more than 40 members from the computer and information industries, applications areas such as medicine and environmental technology, and a variety of government agencies. These members participate in a variety of research and technology transfer projects, and the CIIR supports a number of prototype and operational retrieval applications, such as THOMAS and American Memory at the Library of Congress, InfoSeek, STAT/USA at the Department of Commerce, the Lotus Help Desk, and the U.S. Business Advisor.

The following list describes ten of the most important issues we have encountered during our interactions with CIIR members (apologies to David Letterman). They are listed in approximate reverse order of importance, based purely on my own assessment.

RELEVANCE FEEDBACK

Relevance feedback is a process where users identify relevant documents in an initial list of retrieved documents, and the system then creates a new query based on those sample relevant documents. Algorithms for automatic relevance feedback have been studied in IR for more than thirty years, and the research community considers them to be thoroughly tested and effective. Companies and government agencies that use IR systems also view relevance feedback as a desirable feature, but there are some practical difficulties that have delayed the general adoption of this technique.

Most of the relevance feedback experiments reported in the IR literature were based on small test collections of abstract-length documents. The central problems in relevance feedback are selecting "features" (words, phrases) from relevant documents and calculating weights for these features in the context of a new query. These problems are substantially more difficult in environments with large databases of full-text documents. In addition, people searching databases in real applications often use relevance feedback in different ways than anticipated by IR researchers. Feedback techniques were developed to improve an initial query and assumed that a few relevant documents (all those in the top ten, for example) would be provided. In many real interactions, however, users specify only a single relevant document. Sometimes that relevant document may not even be strongly related to the initial query, and the user is, in effect, browsing using feedback.

These factors mean that traditional feedback techniques can be unpredictable in operational settings. Research aimed at correcting this problem is underway and more operational systems using relevance feedback can be expected in the near future. Relevance feedback techniques are also an important part of building profiles in a routing system (see **routing and filtering**), with the main difference being the number of example relevant documents available.

INFORMATION EXTRACTION

Information extraction techniques, primarily developed in the context of the Advanced Research Projects Agency (ARPA) Message Understanding Conferences (MUCs), are designed to identify database entities, attributes and relationships in full text. For example, for people interested in new joint ventures, an information extraction system could identify the names of the companies involved, the new company, the products, and the location, all from articles coming over a news feed. Companies and government agencies have considerable interest in these techniques, and see them as contributing significant "added-value" to the text databases they and others generate. Potential users also see these techniques as tools to help with data analysis, browsing, and mining using text databases. The current state of information extraction tools is such that it requires a considerable investment to build a new extraction application, and certain types of information are very difficult to identify. Research in this area is focused on reducing the effort required for new applications.

Extraction of simple categories of information is, on the other hand, practical and can be an important part of a text-based information system. Examples of this type of information include company and other organization names, peoples' names, locations, and dates.

MULTIMEDIA RETRIEVAL

Multimedia indexing and retrieval refers to techniques being developed to access image, video and sound databases without text descriptions. The perceived value of multimedia information systems is very high and, consequently, industry has a considerable interest in the development of these techniques. General solutions to multimedia indexing are very difficult and, where they currently exist, tend to be of limited utility. An example of this is indexing images by their color distribution. This technique can be effectively used in some applications, such as retrieving pictures of fabric in specified color shades, but in many other applications simply cannot be used. Some progress has been made in multimedia indexing for specific applications (for example, retrieval of photographs of faces), and in processing language-related multimedia. Examples of language-related multimedia include text in

images, scanned document images, and speech.

Given the number of industrial and academic research groups working in this area, steady improvement of the techniques available can be expected.

EFFECTIVE RETRIEVAL

The development of effective retrieval techniques has been the core of IR research for more than 30 years. A number of measures of effectiveness have been proposed, but the most frequently mentioned are recall and precision. Finding text that satisfies a user's information need is not simple, and considerable progress has been made in developing ranking techniques that are significantly more effective than Boolean logic.

Contrary to some researchers' opinions, companies that sell and use IR systems *are* interested in effectiveness. Having a more effective retrieval engine is a major selling point. It is not, however, the primary focus of their concerns and I have indicated this by the quite low ranking of this issue in the top 10. With regard to effectiveness, companies are particularly interested in techniques that produce significant improvements (rather than a few percent average precision) and that avoid occasional major mistakes. A system that performs well on most queries but makes it difficult for users to recover from bad mistakes, or even understand why they were made, is likely to be considered unacceptable. These occasional mistakes have very little impact on the average recall/precision measures used in standard IR tests, but considerable impact on end users. An example of a technique that produces reliable (although small) improvements in effectiveness, is generally well-regarded by users, but is one of the main sources of occasional bad mistakes is stemming. Solutions include building better stemmers and doing stemming as part of query processing rather than indexing.

ROUTING AND FILTERING

Information routing, filtering and clipping are all synonyms used to describe the process of identifying relevant documents in streams of information such as news feeds. Instead of comparing a single query to large numbers of archived documents, as is the case for IR, large number of archived profiles are compared to individual documents. Documents that match are sent to the users associated with the profile. A profile is a representation of a long-term information need and is usually more complex than a session-based query.

Companies and government agencies often indicate that routing is the main function required for a text-based system, with IR being a backup, secondary function. Given that, we can expect to see a significant increase in the use of routing systems.

Both efficiency and effectiveness are important for routing, and both have to be addressed in different ways than IR systems. Efficiency is important in order to deal with high-volume document streams (e.g. 100 MB/hour) and large numbers of profiles (tens of thousands). New indexing and memory-based architectures are being developed for these systems. In terms of effectiveness, the basic algorithms are very similar to retrieval but instead of producing a ranking, cutoffs must be used to separate relevant from non-relevant documents. Learning techniques are being studied both as a means of determining these cutoffs and automatically building profiles based on user feedback.

INTERFACES AND BROWSING

Effective interfaces for text-based information systems are a high priority for users of these systems. The interface is a major part of how a system is evaluated, and as the retrieval and routing algorithms become more complex to improve recall and precision, more stress is placed on the design of interfaces that make the system easy to use and understandable. Interfaces must support a range of functions including query formulation, presentation of retrieved information, feedback, and browsing. The challenge is present this sophisticated functionality in a conceptually simple way. Despite the importance of this issue, there has been relatively little relevant research done by either the IR or

human-computer interface (HCI) communities. This, however, is changing and much more work on interfaces for information systems and information visualization will be appearing.

"MAGIC"

One of the major causes of failures in IR systems is vocabulary mismatch. This means that the information need is often described using different words than are found in relevant documents. Techniques that address this problem by automatic expansion of the query are often regarded as a form of "magic" by users and are viewed as highly desirable. Vocabulary expansion can result from transforming the document and query representations, as with Latent Semantic Indexing, or it can be done as a form of automatic thesaurus built by corpus analysis. Further research in this area will make these techniques more reliable and efficient.

EFFICIENT, FLEXIBLE INDEXING AND RETRIEVAL

One of the most frequently mentioned, and most highly rated, issues is efficiency. Many different aspects of a system can have an impact on efficiency, and metrics such as query response time and indexing speed are major concerns of virtually every company involved with text-based systems. In the past, efficiency was a secondary issue in much of the IR literature. This has changed with the increased accessibility of large, full-text databases and new algorithms designed to increase indexing and query speed are published regularly. There has also been substantial research on text compression algorithms for decreasing storage overheads and I/O times. New applications of text-based systems that involve real-time and multi-user constraints will also require more work on concurrency control, update, and recovery strategies appropriate for these applications. Web-based systems, such as Infoseek, have particularly strict efficiency requirements since they must deal with hundreds of thousands of queries per day.

The other aspect of indexing that is considered very important is the capability of handling a wide variety of document formats. This includes both standards such as SGML, HTML, Acrobat, and WordPerfect to name a few, as well as the myriad formats used in text-based applications such as medicine, law, commerce, etc.

DISTRIBUTED IR

With the advent of the World-Wide Web and the huge increase in the use of the Internet, there has been a corresponding increase in demand for text retrieval systems that can work in distributed, wide-area network environments. This demand also comes from groupware applications such as Lotus Notes, which facilitate the rapid creation of databases distributed throughout an organization. One problem of this type is addressed by Web search engines, such as Infoseek and Lycos, which index Web pages and provide access to them. The more general problems are locating the best databases to search in a distributed environment that may contain hundreds or even thousands of databases, and merging the results that come back from the distributed search. The results must be merged in order to produce the overall ranking of retrieved items, instead of a collection of individual rankings. The difficulty of doing this effectively comes from the fact that the individual rankings may be incompatible in the sense that the numbers used to produce these rankings may not be directly comparable (they may even come from different IR systems). Research addressing these issues has begun to appear in the major conferences.

The work in this area, and that done in routing and relevance feedback, is closely related to the work on multi-agent systems that has received a lot of attention.

INTEGRATED SOLUTIONS

The most important problem from the point of view of companies using and selling text-based systems is integration with other systems. A text retrieval system is a tool that can be used to solve part of an organization's information management problems. It is not often, however, the complete solution.

Typically, a complete solution requires other text-based tools such as routing and extraction, tools for handling multimedia and scanned documents such as OCR, a database management system for structured data, and workflow or other groupware systems for managing documents and their use in the organization. Currently, these systems must be integrated using customized software, and even then the integration is often very superficial. More work on standardized architectures and common platforms is needed. Examples of some of the standards efforts include the Z39.50 search protocol, the SQL-Multimedia proposals, and the ARPA TIPSTER architecture for integrating retrieval, routing, and extraction systems.

One of the most important aspects of developing common platforms is the integration of database management and IR systems. An effective integration of these systems, together with multimedia capabilities, would provide an information system that could be used to manage many of the applications that currently exist. Although partial solutions to this integration do exist, they do not address one of the fundamental issues, which is that database systems all retrieve using Boolean logic, whereas both text and multimedia require techniques involving uncertainty and ranking for effective retrieval. True integration of text and multimedia is likely to require significant changes in the standard database techniques for indexing and query optimization, and may even require new query languages. Research on these issues is underway and more effort and support for this work is likely.

This list of ten issues is not meant to be comprehensive. There are other research areas that are of significant interest to companies, such as multilingual IR, data mining in text databases, and text categorization (attaching categories to text from predefined set, such as news categories or diagnosis codes). Although the ranking of these issues may not coincide with the priorities that IR researchers (including myself) may put on their own research programs, there is almost total agreement that these issues all pose significant research challenges. Some of these challenges will be solved in the short-term, but others will be the basis of longer research projects. The opportunities and experience provided by the explosion of operational text-based systems will be invaluable for IR research.

The CIIR Bibliography describes papers that have more details on the related research being done at the CIIR, as well as pointers to other peoples' work.

CROFT. This paper was first published in *D-LIB Magazine*, November 1995. It has been reproduced here with permission from the publishers